Forecasting economic variables with nonlinear models

Timo Teräsvirta
Department of Economic Statistics
Stockholm School of Economics
Box 6501, SE-113 83 Stockholm, Sweden

SSE/EFI Working Paper Series in Economics and Finance
No. 598

December 29, 2005

Abstract

This chapter is concerned with forecasting from nonlinear conditional mean models. First, a number of often applied nonlinear conditional mean models are introduced and their main properties discussed. The next section is devoted to techniques of building nonlinear models. Ways of computing multi-step-ahead forecasts from nonlinear models are surveyed. Tests of forecast accuracy in the case where the models generating the forecasts may be nested are discussed. There is a numerical example, showing that even when a stationary nonlinear process generates the observations, future observations may in some situations be better forecast by a linear model with a unit root. Finally, some empirical studies that compare forecasts from linear and nonlinear models are discussed.

JEL Classification Codes. C22, C45, C53

Keywords. Forecast accuracy; forecast comparison; hidden Markov model; neural network; nonlinear modelling; recursive forecast; smooth transition regression; switching regression

Acknowledgements. Financial support from Jan Wallander's and Tom Hedelius's Foundation, Grant No. J02-35, is gratefully acknowledged. Discussions with Clive Granger have been very helpful. I also wish to thank three anonymous referees, Marcelo Medeiros and Dick van Dijk for useful comments but retain responsibility for any errors and shortcomings in this work.

1 Introduction

In recent years, nonlinear models have become more common in empirical economics than they were a few decades ago. This trend has brought with it an increased interest in forecasting economic variables with nonlinear models: for recent accounts of this topic, see Tsay (2002) and Clements, Franses and Swanson (2004). Nonlinear forecasting has also been discussed in books on nonlinear economic modelling such as Granger and Teräsvirta (1993, Chapter 9) and Franses and van Dijk (2000). More specific surveys include Zhang, Patuwo and Hu (1998) on forecasting (not only economic forecasting) with neural network models and Lundbergh and Teräsvirta (2002) who consider forecasting with smooth transition autoregressive models. Ramsey (1996) discusses difficulties in forecasting economic variables with nonlinear models. Large-scale comparisons of the forecasting performance of linear and nonlinear models have appeared in the literature; see Stock and Watson (1999), Marcellino (2002) and Teräsvirta, van Dijk and Medeiros (2005) for examples. There is also a growing literature consisting of forecast comparisons that involve a rather limited number of time series and nonlinear models as well as comparisons entirely based on simulated series.

There exist an unlimited amount of nonlinear models, and it is not possible to cover all developments in this survey. The considerations are restricted to parametric nonlinear models, which excludes forecasting with nonparametric models. For information on nonparametric forecasting, the reader is referred to Fan and Yao (2003). Besides, only a small number of frequently applied parametric nonlinear models are discussed here. It is also worth mentioning that the interest is solely focussed on stochastic models. This excludes deterministic processes such as chaotic ones. This is motivated by the fact that chaos is a less useful concept in economics than it is in natural sciences. Another area of forecasting with nonlinear models that is not covered here is volatility forecasting. The reader is referred to Andersen, Bollerslev and Christoffersen (2006) and the survey by Poon and Granger (2003).

The plan of the chapter is the following. In Section 2, a number of parametric nonlinear models are presented and their properties briefly discussed. Section 3 is devoted to strategies of building certain types of nonlinear models. In Section 4 the focus shifts to forecasting, more specifically, to different methods of obtaining multistep forecasts. Combining forecasts is also briefly mentioned. Problems in and ways of comparing the accuracy of point forecasts from linear and nonlinear models is considered in Section 5, and a specific simulated example of such a comparison in Section 6. Empirical forecast comparisons form the topic of Section 7, and Section 8 contains final remarks.

2 Nonlinear models

2.1 General

Regime-switching has been a popular idea in economic applications of nonlinear models. The data-generating process to be modelled is perceived as a linear process that switches between a number of regimes according to some rule. For example, it may be argued that the dynamic properties of the growth rate of the volume of industrial production or gross national product process are different in recessions and expansions. As another example, changes in government policy may instigate switches in regime.

These two examples are different in nature. In the former case, it may be assumed that nonlinearity is in fact controlled by an observable variable such as a lag of the growth rate. In the latter one, an observable indicator for regime switches may not exist. This feature will lead to a family of nonlinear models different from the previous one.

In this chapter we present a small number of special cases of the nonlinear dynamic regression model. These are rather general models in the sense that they have not been designed for testing a particular economic theory proposition or describing economic behaviour in a particular situation. They share this property with the dynamic linear model. No clear-cut rules for choosing a particular nonlinear family exist, but the previous examples suggest that in some cases, choices may be made *a priori*. Estimated models can, however, be compared *ex post*. In theory, nonnested tests offer such a possibility, but applying them in the nonlinear context is more demanding that in the linear framework, and few, if any, examples of that exist in the literature. Model selection criteria are sometimes used for the purpose as well as post-sample forecasting comparisons. It appears that successful model building, that is, a systematic search to find a model that fits the data well, is only possible within a well-defined family of nonlinear models. The family of autoregressive $-$ moving average models constitutes a classic linear example; see Box and Jenkins (1970). Nonlinear model building is discussed in Section 3.

2.2 Nonlinear dynamic regression model

A general nonlinear dynamic model with an additive noise component can be defined as follows:

$$y_t = f(\mathbf{z}_t; \theta) + \varepsilon_t \tag{1}$$

where $\mathbf{z}_t = (\mathbf{w}_t', \mathbf{x}_t')'$ is a vector of explanatory variables, $\mathbf{w}_t = (1, y_{t-1}, ..., y_{t-p})'$, and the vector of strongly exogenous variables $\mathbf{x}_t = (x_{1t}, ..., x_{kt})'$. Furthermore, $\varepsilon_t \sim \text{iid}(0, \sigma^2)$. It is assumed that y_t is a stationary process. Nonsta-

tionary nonlinear processes will not be considered in this survey. Many of the models discussed in this section are special cases of (1) that have been popular in forecasting applications. Moving average models and models with stochastic coefficients, an example of so-called doubly stochastic models, will also be briefly highlighted.

Strict stationarity of (1) may be investigated using the theory of Markov chains. Tong (1990, Chapter 4) contains a discussion of the relevant theory. Under a condition concerning the starting distribution, geometric ergodicity of a Markov chain implies strict stationarity of the same chain, and a set of conditions for geometric ergodicity are given. These results can be used for investigating strict stationarity in special cases of (1) , as the model can be expressed as a $(p+1)$-dimensional Markov chain. As an example (Example 4.3 in Tong, 1990), consider the following modification of the exponential smooth transition autoregressive (ESTAR) model to be discussed in the next section:

$$\begin{aligned} y_t &= \sum_{j=1}^{p}[\phi_j y_{t-j} + \theta_j y_{t-j}(1 - \exp\{-\gamma y_{t-j}^2\})] + \varepsilon_t \\ &= \sum_{j=1}^{p}[(\phi_j + \theta_j) y_{t-j} - \theta_j y_{t-j} \exp\{-\gamma y_{t-j}^2\}] + \varepsilon_t \end{aligned} \tag{2}$$

where $\{\varepsilon_t\} \sim \text{iid}(0, \sigma^2)$. It can be shown that (2) is geometrically ergodic if the roots of $1 - \sum_{j=1}^{p}(\phi_j + \theta_j)L^j$ lie outside the unit circle. This result partly relies on the additive structure of this model. In fact, it is not known whether the same condition holds for the following, more common but non-additive, ESTAR model:

$$y_t = \sum_{j=1}^{p}[\phi_j y_{t-j} + \theta_j y_{t-j}(1 - \exp\{-\gamma y_{t-d}^2\})] + \varepsilon_t, \gamma > 0$$

where $d > 0$ and $p > 1$.

As another example, consider the first-order self-exciting threshold autoregressive (SETAR) model (see Section 2.4)

$$y_t = \phi_{11} y_{t-1} I(y_{t-1} \leq c) + \phi_{12} y_{t-1} I(y_{t-1} > c) + \varepsilon_t$$

where $I(A)$ is an indicator function: $I(A) = 1$ when event A occurs; zero otherwise. A necessary and sufficient condition for this SETAR process to be geometrically ergodic is $\phi_{11} < 1$, $\phi_{12} < 1$ and $\phi_{11}\phi_{12} < 1$. For higher-order models, normally only sufficient conditions exist, and for many interesting models these conditions are quite restrictive. An example will be given in Section 2.4.

2.3 Smooth transition regression model

The smooth transition regression (STR) model originated in the work of Bacon and Watts (1971). These authors considered two regression lines and devised a model in which the transition from one line to the other is smooth. They used the hyperbolic tangent function to characterize the transition. This function is close to both the normal cumulative distribution function and the logistic function. Maddala (1977, p. 396) in fact recommended the use of the logistic function as transition function, and this has become the prevailing standard; see, for example, Teräsvirta (1998). In general terms we can define the STR model as follows:

$$\begin{aligned} y_t &= \phi'\mathbf{z}_t + \theta'\mathbf{z}_t G(\gamma, \mathbf{c}, s_t) + \varepsilon_t \\ &= \{\phi + \theta G(\gamma, \mathbf{c}, s_t)\}'\mathbf{z}_t + \varepsilon_t, t = 1, ..., T \end{aligned} \tag{3}$$

where $\mathbf{z}_t$ is defined as in (1), $\phi = (\phi_0, \phi_1, ..., \phi_m)'$ and $\theta = (\theta_0, \theta_1, ..., \theta_m)'$ are parameter vectors, and $\varepsilon_t \sim \text{iid}(0, \sigma^2)$. In the transition function $G(\gamma, \mathbf{c}, s_t)$, γ is the slope parameter and $\mathbf{c} = (c_1, ..., c_K)'$ a vector of location parameters, $c_1 \leq ... \leq c_K$. The transition function is a bounded function of the transition variable s_t, continuous everywhere in the parameter space for any value of s_t. The last expression in (3) indicates that the model can be interpreted as a linear model with stochastic time-varying coefficients $\phi + \theta G(\gamma, \mathbf{c}, s_t)$ where s_t controls the time-variation. The logistic transition function has the general form

$$G(\gamma, \mathbf{c}, s_t) = (1 + \exp\{-\gamma \prod_{k=1}^{K}(s_t - c_k)\})^{-1}, \gamma > 0 \tag{4}$$

where $\gamma > 0$ is an identifying restriction. Equation (3) jointly with (4) defines the logistic STR (LSTR) model. The most common choices for K are $K = 1$ and $K = 2$. For $K = 1$, the parameters $\phi + \theta G(\gamma, \mathbf{c}, s_t)$ change monotonically as a function of s_t from ϕ to $\phi + \theta$. For $K = 2$, they change symmetrically around the mid-point $(c_1 + c_2)/2$ where this logistic function attains its minimum value. The minimum lies between zero and 1/2. It reaches zero when $\gamma \to \infty$ and equals 1/2 when $c_1 = c_2$ and $\gamma < \infty$. Slope parameter γ controls the slope and c_1 and c_2 the location of the transition function.

The LSTR model with $K = 1$ (LSTR1 model) is capable of characterizing asymmetric behaviour. As an example, suppose that s_t measures the phase of the business cycle. Then the LSTR1 model can describe processes whose dynamic properties are different in expansions from what they are in recessions, and the transition from one extreme regime to the other is smooth.

The LSTR2 model is appropriate in situations where the local dynamic behaviour of the process is similar at both large and small values of s_t and different in the middle.

When $\gamma = 0$, the transition function $G(\gamma, \mathbf{c}, s_t) \equiv 1/2$ so that STR model (3) nests a linear model. At the other end, when $\gamma \to \infty$ the LSTR1 model approaches the switching regression (SR) model, see Section 2.4, with two regimes and $\sigma_1^2 = \sigma_2^2$. When $\gamma \to \infty$ in the LSTR2 model, the result is a switching regression model with three regimes such that the outer regimes are identical and the mid-regime different from the other two.

Another variant of the LSTR2 model is the exponential STR (ESTR, in the univariate case ESTAR) model in which the transition function

$$G(\gamma, c, s_t) = 1 - \exp\{-\gamma(s_t - c)^2\}, \gamma > 0 \tag{5}$$

This transition function is an approximation to (4) with $K = 2$ and $c_1 = c_2$. When $\gamma \to \infty$, however, $G(\gamma, c, s_t) = 1$ for $s_t \neq c$, in which case equation (3) is linear except at a single point. Equation (3) with (5) has been a popular tool in investigations of the validity of the purchasing power parity (PPP) hypothesis; see for example the survey by Taylor and Sarno (2002).

In practice, the transition variable s_t is a stochastic variable and very often an element of $\mathbf{z}_t$. It can also be a linear combination of several variables. A special case, $s_t = t$, yields a linear model with deterministically changing parameters. Such a model has a role to play, among other things, in testing parameter constancy, see Section 2.7.

When $\mathbf{x}_t$ is absent from (3) and $s_t = y_{t-d}$ or $s_t = \Delta y_{t-d}$, $d > 0$, the STR model becomes a univariate smooth transition autoregressive (STAR) model. The logistic STAR (LSTAR) model was introduced in the time series literature by Chan and Tong (1986) who used the density of the normal distribution as the transition function. The exponential STAR (ESTAR) model appeared already in Haggan and Ozaki (1981). Later, Teräsvirta (1994) defined a family of STAR models that included both the LSTAR and the ESTAR model and devised a data-driven modelling strategy with the aim of, among other things, helping the user to choose between these two alternatives.

Investigating the PPP hypothesis is just one of many applications of the STR and STAR models to economic data. Univariate STAR models have been frequently applied in modelling asymmetric behaviour of macroeconomic variables such as industrial production and unemployment rate, or nonlinear behaviour of inflation. In fact, many different nonlinear models have been fitted to unemployment rates; see Proietti (2003) for references. As to STR models, several examples of the its use in modelling money demand such as Teräsvirta and Eliasson (2001) can be found in the literature.

Venetis, Paya and Peel (2003) recently applied the model to a much investigated topic: usefulness of the interest rate spread in predicting output growth. The list of applications could be made longer.

2.4 Switching regression and threshold autoregressive model

The standard switching regression model is piecewise linear, and it is defined as follows:

$$y_t = \sum_{j=1}^{r+1} (\phi_j' \mathbf{z}_t + \varepsilon_{jt}) I(c_{j-1} < s_t \leq c_j) \tag{6}$$

where $\mathbf{z}_t = (\mathbf{w}_t', \mathbf{x}_t')'$ is defined as before, s_t is a switching variable, usually assumed to be a continuous random variable, $c_0, c_1, ..., c_{r+1}$ are threshold parameters, $c_0 = -\infty$, $c_{r+1} = +\infty$. Furthermore, $\varepsilon_{jt} \sim \text{iid}(0, \sigma_j^2)$, $j = 1, ..., r$. It is seen that (6) is a piecewise linear model whose switch-points, however, are generally unknown. A popular alternative in practice is the two-regime SR model

$$y_t = (\phi_1' \mathbf{z}_t + \varepsilon_{1t}) I(s_t \leq c_1) + (\phi_2' \mathbf{z}_t + \varepsilon_{2t}) \{1 - I(s_t \leq c_1)\}. \tag{7}$$

It is a special case of the STR model (3) with $K = 1$ in (4).

When $\mathbf{x}_t$ is absent and $s_t = y_{t-d}, d > 0$, (6) becomes the self-exciting threshold autoregressive (SETAR) model. The SETAR model has been widely applied in economics. A comprehensive account of the model and its statistical properties can be found in Tong (1990). A two-regime SETAR model is a special case of the LSTAR1 model when the slope parameter $\gamma \to \infty$.

A special case of the SETAR model itself, suggested by Enders and Granger (1998) and called the momentum-TAR model, is the one with two regimes and $s_t = \Delta y_{t-d}$. This model may be used to characterize processes in which the asymmetry lies in growth rates: as an example, the growth of the series when it occurs may be rapid but the return to a lower level slow.

It was mentioned in Section 2.2 that stationarity conditions for higher-order models can often be quite restrictive. As an example, consider the univariate SETAR model of order p, that is, $\mathbf{x}_t \equiv \mathbf{0}$ and $\phi_j = (1, \phi_{j1}, ..., \phi_{jp})'$ in (6). Chan (1993) contains a sufficient condition for this model to be stationary. It has the form

$$\max_i \sum_{j=1}^{p} |\phi_{ji}| < 1.$$

For $p = 1$ the condition becomes $\max_i |\phi_{1i}| < 1$, which is already in this simple case a more restrictive condition than the necessary and sufficient condition presented in Section 2.2.

The SETAR model has also been a popular tool in investigating the PPP hypothesis; see the survey by Taylor and Sarno (2002). Like the STAR model, the SETAR model has been widely applied to modelling asymmetries in macroeconomic series. It is often argued that the US interest rate processes have more than one regime, and SETAR models have been fitted to these series, see Pfann, Schotman and Tschernig (1996) for an example. These models have also been applied to modelling exchange rates as in Henry, Olekalns and Summers (2001) who were, among other things, interested in the effect of the East-Asian 1997-1998 currency crisis on the Australian dollar.

2.5 Markov-switching model

In the switching regression model (6), the switching variable is an observable continuous variable. It may also be an unobservable variable that obtains a finite number of discrete values and is independent of y_t at all lags, as in Lindgren (1978). Such a model may be called the Markov-switching or hidden Markov regression model, and it is defined by the following equation:

$$y_t = \sum_{j=1}^{r} \alpha_j' \mathbf{z}_t I(s_t = j) + \varepsilon_t \tag{8}$$

where $\{s_t\}$ follows a Markov chain, often of order one. If the order equals one, the conditional probability of the event $s_t = i$ given s_{t-k}, $k = 1, 2, ...$, is only dependent on s_{t-1} and equals

$$\Pr\{s_t = i | s_{t-1} = j\} = p_{ij},\ i, j = 1, ..., r \tag{9}$$

such that $\sum_{i=1}^{r} p_{ij} = 1$. The transition probabilities p_{ij} are unknown and have to be estimated from the data. The error process ε_t is often assumed not to be dependent on the 'regime' or the value of s_t, but the model may be generalized to incorporate that possibility. In its univariate form, $\mathbf{z}_t = \mathbf{w}_t$, model (8) with transition probabilities (9) has been called the suddenly changing autoregressive (SCAR) model; see Tyssedal and Tjøstheim (1988).

There is a Markov-switching autoregressive model, proposed by Hamilton (1989), that is more common in econometric applications than the SCAR model. In this model, the intercept is time-varying and determined by the

value of the latent variable s_t and its lags. It has the form

$$y_t = \mu_{s_t} + \sum_{j=1}^{p} \alpha_j (y_{t-j} - \mu_{s_{t-j}}) + \varepsilon_t \tag{10}$$

where the behaviour of s_t is defined by (9) , and $\mu_{s_t} = \mu^{(i)}$ for $s_t = i$, such that $\mu^{(i)} \neq \mu^{(j)}$, $i \neq j$. For identification reasons, y_{t-j} and $\mu_{s_{t-j}}$ in (10) share the same coefficient. The stochastic intercept of this model, $\mu_{s_t} - \sum_{j=1}^{p} \alpha_j \mu_{s_{t-j}}$, thus can obtain r^{p+1} different values, and this gives the model the desired flexibility. A comprehensive discussion of Markov-switching models can be found in Hamilton (1994, Chapter 22).

Markov-switching models can be applied when the data can be conveniently thought of as having been generated by a model with different regimes such that the regime changes do not have an observable or quantifiable cause. They may also be used when data on the switching variable is not available and no suitable proxy can be found. This is one of the reasons why Markov-switching models have been fitted to interest rate series, where changes in monetary policy have been a motivation for adopting this approach. Modelling asymmetries in macroeconomic series has, as in the case of SETAR and STAR models, been another area of application; see Hamilton (1989) who fitted a Markov-switching model of type (10) to the post World War II quarterly US GNP series. Tyssedal and Tjøstheim (1988) fitted a three-regime SCAR model to a daily IBM stock return series originally analyzed in Box and Jenkins (1970).

2.6 Autoregressive neural network model

Modelling various processes and phenomena, including economic ones, using artificial neural network (ANN) models has become quite popular. Many textbooks have been written about these models, see, for example, Fine (1999) or Haykin (1999). A detailed treatment can be found in White (2006), whereas the discussion here is restricted to the simplest single-equation case, which is the so-called "single hidden-layer" model. It has the following form:

$$y_t = \beta_0' \mathbf{z}_t + \sum_{j=1}^{q} \beta_j G(\gamma_j' \mathbf{z}_t) + \varepsilon_t \tag{11}$$

where y_t is the output series, $\mathbf{z}_t = (1, y_{t-1}, ..., y_{t-p}, x_{1t}, ..., x_{kt})'$ is the vector of inputs, including the intercept and lagged values of the output, $\beta_0' \mathbf{z}_t$ is a linear unit, and $\beta_j, j = 1, ..., q$, are parameters, called "connection strengths" in the neural network literature. Many neural network modellers exclude the

linear unit altogether, but it is a useful component in time series applications. Furthermore, function $G(.)$ is a bounded function called "the squashing function" and γ_j, $j = 1, ..., q$, are parameter vectors. Typical squashing functions are monotonically increasing ones such as the logistic function and the hyperbolic tangent function and thus have the same form as transition functions of STAR models. The so-called radial basis functions that resemble density functions are another possibility. The errors ε_t are often assumed iid$(0,\sigma^2)$. The term "hidden layer" refers to the structure of (11). While the output y_t and the input vector $\mathbf{z}_t$ are observed, the linear combination $\sum_{j=1}^{q} \beta_j G(\gamma_j' \mathbf{z}_t)$ is not. It thus forms a hidden layer between the "output layer" y_t and "input layer" $\mathbf{z}_t$.

A theoretical argument used to motivate the use of ANN models is that they are universal approximators. Suppose that $y_t = H(\mathbf{z}_t)$, that is, there exists a functional relationship between y_t and $\mathbf{z}_t$. Then, under mild regularity conditions for H, there exists a positive integer $q \leq q_0 < \infty$ such that for an arbitrary $\delta > 0$, $|H(\mathbf{z}_t) - \sum_{j=1}^{q} \beta_j G(\gamma_j' \mathbf{z}_t)| < \delta$. The importance of this result lies in the fact that q is finite, whereby any unknown function H can be approximated arbitrarily accurately by a linear combination of squashing functions $G(\gamma_j' \mathbf{z}_t)$. This has been discussed in several papers including Cybenko (1989), Funahashi (1989), Hornik, Stinchombe and White (1989) and White (1990).

A statistical property separating the artificial neural network model (11) from other nonlinear econometric models presented here is that it is only locally identified. It is seen from equation (11) that the hidden units are exchangeable. For example, letting any $(\beta_i, \gamma_i')'$ and $(\beta_j, \gamma_j')', i \neq j$, change places in the equation does not affect the value of the likelihood function. Thus for $q > 1$ there always exists more than one observationally equivalent parameterization, so that additional parameter restrictions are required for global identification. Furthermore, the sign of one element in each γ_j, the first one, say, has to be fixed in advance to exclude observationally equivalent parameterizations. The identification restrictions are discussed, for example, in Hwang and Ding (1997).

The rich parameterization of ANN models makes the estimation of parameters difficult. Computationally feasible, yet effective, shortcuts are proposed and implemented in White (2006). Goffe, Ferrier and Rogers (1994) contains an example showing that simulated annealing, which is a heuristic estimation method, may be a powerful tool in estimating parameters of these models. ANN models have been fitted to various economic time series. Since the model is a universal approximator rather than one with parameters with economic interpretation, the purpose of fitting these models has mainly been forecasting. Examples of their performance in forecasting macroeconomic

variables can be found in Section 7.3.

2.7 Time-varying autoregressive model

A time-varying regression model is an STR model in which the transition variable $s_t = t$. It can thus be defined as follows:

$$y_t = \phi' \mathbf{z}_t + \theta' \mathbf{z}_t G(\gamma, \mathbf{c}, t) + \varepsilon_t, \ t = 1, ..., T \tag{12}$$

where the transition function

$$G(\gamma, \mathbf{c}, s_t) = (1 + \exp\{-\gamma \prod_{k=1}^{K} (t - c_k)\})^{-1}, \gamma > 0. \tag{13}$$

When $K = 1$ and $\gamma \to \infty$ in (13) , equation (12) represents a linear dynamic regression model with a break in parameters at $t = c_1$. It can be generalized to a model with several transitions:

$$y_t = \phi' \mathbf{z}_t + \sum_{j=1}^{r} \theta_j' \mathbf{z}_t G_j(\gamma_j, \mathbf{c}_j, t) + \varepsilon_t, \ t = 1, ..., T \tag{14}$$

where transition functions G_j typically have the form (13) with $K = 1$. When $\gamma_j \to \infty$, $j = 1, ..., r$, in (14) , the model becomes a linear model with multiple breaks. Specifying such models has recently received plenty of attention; see, for example, Bai and Perron (1998, 2003) and Banerjee and Urga (2005). In principle, these models should be preferable to linear models without breaks because the forecasts are generated from the most recent specification instead of an average one, which is the case if the breaks are ignored. In practice, the number of break-points and their locations have to be estimated from the data, which makes this suggestion less straightforward. Even if this difficulty is ignored, it may be optimal to use pre-break observations in forecasting. The reason is that while the one-step-ahead forecast based on post-break data is unbiased (if the model is correctly specified), it may have a large variance. The mean square error of the forecast may be reduced if the model is estimated by using at least some pre-break observations as well. This introduces bias but at the same time reduces the variance. For more information of this bias-variance tradeoff, see Pesaran and Timmermann (2002).

Time-varying coefficients can also be stochastic:

$$y_t = \phi_t' \mathbf{z}_t + \varepsilon_t, \ t = 1, ..., T \tag{15}$$

where $\{\phi_t\}$ is a sequence of random variables. In a large forecasting study, Marcellino (2002) assumed that $\{\phi_t\}$ was a random walk, that is, $\{\Delta\phi_t\}$ was

a sequence of normal independent variables with zero mean and a known variance. This assumption is a testable alternative to parameter constancy; see Nyblom (1989). For the estimation of stochastic random coefficient models, the reader is referred to Harvey (2006). Another assumption, albeit a less popular one in practice, is that $\{\phi_t\}$ follows a stationary vector autoregressive model. Parameter constancy in (15) may be tested against this alternative as well: see Watson and Engle (1985) and Lin and Teräsvirta (1999).

2.8 Nonlinear moving average models

Nonlinear autoregressive models have been quite popular among practitioners, but nonlinear moving average models have also been proposed in the literature. A rather general nonlinear moving average model of order q may be defined as follows:

$$y_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-q}; \theta) + \varepsilon_t$$

where $\{\varepsilon_t\} \sim \text{iid}(0, \sigma^2)$. A problem with these models is that their invertibility conditions may not be known, in which case the models cannot be used for forecasting. A common property of moving average models is that if the model is invertible, forecasts from it for more than q steps ahead equal the unconditional mean of y_t. Some nonlinear moving average models are linear in parameters, which makes forecasting with them easy in the sense that no numerical techniques are required when forecasting several steps ahead. As an example of a nonlinear moving average model, consider the asymmetric moving average (asMA) model of Wecker (1981). It has the form

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{j=1}^{q} \psi_j I(\varepsilon_{t-j} > 0)\varepsilon_{t-j} + \varepsilon_t \tag{16}$$

where $I(\varepsilon_{t-j} > 0) = 1$ when $\varepsilon_{t-j} > 0$ and zero otherwise, and $\{\varepsilon_t\} \sim \text{nid}(0, \sigma^2)$. This model has the property that the effects of a positive shock and a negative shock of the same sizes on y_t are not symmetric when $\psi_j \neq 0$ for at least one j, $j = 1, ..., q$.

Brännäs and De Gooijer (1994) extended (16) to contain a linear autoregressive part and called the model an autoregressive asymmetric moving average (ARasMA) model. The forecasts from an ARasMA model has the property that after q steps ahead they are identical to the forecasts from a linear AR model that has the same autoregressive parameters as the ARasMA model. This implies that the forecast densities more than q periods ahead are symmetric, unless the error distribution is asymmetric.

3 Building nonlinear models

Building nonlinear models comprises three stages. First, the structure of the model is specified, second, its parameters are estimated and third, the estimated model has to be evaluated before it is used for forecasting. The last stage is important because if the model does not satisfy in-sample evaluation criteria, it cannot be expected to produce accurate forecasts. Of course, good in-sample behaviour of a model is not synonymous with accurate forecasts, but in many cases it may at least be viewed as a necessary condition for obtaining such forecasts from the final model.

It may be argued, however, that the role of model building in constructing models for forecasting is diminishing because computations has become inexpensive. It is easy to estimate a possibly large number of models and combine the forecasts from them. This suggestion is related to thick modelling that Granger and Jeon (2004) recently discussed. A study where this has been a successful strategy will be discussed in Section 7.3.1. On the other hand, many popular nonlinear models such as the smooth transition or threshold autoregressive, or Markov switching models, nest a linear model and are unidentified if the data-generating process is linear. Fitting one of these models to linear series leads to inconsistent parameter estimates, and forecasts from the estimated model are bound to be bad. Combining these forecasts with others would not be a good idea. Testing linearity first, as a part of the modelling process, greatly reduces the probability of this alternative. Aspects of building smooth transition, threshold autoregressive, and Markov switching models will be briefly discussed below.

3.1 Testing linearity

Since many of the nonlinear models considered in this chapter nest a linear model, a short review of linearity testing may be useful. In order to illustrate the identification problem, consider the following nonlinear model:

$$y_t = \phi'\mathbf{z}_t + \theta'\mathbf{z}_t G(\gamma;\mathbf{s}_t) + \varepsilon_t = (\phi + \theta G(\gamma;\mathbf{s}_t))'\mathbf{z}_t + \varepsilon_t \qquad (17)$$

where $\mathbf{z}_t = (1, \widetilde{\mathbf{z}}_t')'$ is an $(m \times 1)$ vector of explanatory variables, some of which can be lags of y_t, and $\{\varepsilon_t\}$ is a white noise sequence with zero mean and $\mathsf{E}\varepsilon_t^2 = \sigma^2$. Depending on the definitions of $G(\gamma;\mathbf{s}_t)$ and $\mathbf{s}_t$, (17) can represent an STR (STAR), SR (SETAR) or a Markov-switching model. The model is linear when $\theta = \mathbf{0}$. When this is the case, parameter vector γ is not identified. It can take any value without the likelihood of the process being affected. Thus, estimating ϕ, θ and γ consistently from (17) is not possible and for this reason, the standard asymptotic theory is not available.

The problem of testing a null hypothesis when the model is only identified under the alternative was first considered by Davies (1977). The general idea is the following. As discussed above, the model is identified when γ is known, and testing linearity of (17) is straightforward. Let $S_T(\gamma)$ be the corresponding test statistic whose large values are critical and define $\Gamma = \{\gamma : \gamma \in \Gamma\}$, the set of admissible values of γ. When γ is unknown, the statistic is not operational because it is a function of γ. Davies (1977) suggested that the problem be solved by defining another statistic $S_T = \sup_{\gamma \in \mathbf{\Gamma}} S_T(\gamma)$ that is no longer a function of γ. Its asymptotic null distribution does not generally have an analytic form, but Davies (1977) gives an approximation to it that holds under certain conditions, including the assumption that $S(\gamma) = \text{plim}_{T\to\infty} S_T(\gamma)$ has a derivative. This, however, is not the case in SR and SETAR models. Other choices of test statistic include the average:

$$S_T = \text{ave} S_T(\gamma) = \int_\Gamma S_T(\gamma) \mathrm{d}W(\gamma) \tag{18}$$

where $W(\gamma)$ is a weight function defined by the user such that $\int_\Gamma W(\gamma)\mathrm{d}\gamma = 1$. Another choice is the exponential:

$$\exp S_T = \ln(\int_\Gamma \exp\{(1/2)S_T(\gamma)\}\mathrm{d}W(\gamma)). \tag{19}$$

see Andrews and Ploberger (1994).

Hansen (1996) shows how to obtain asymptotic critical values for these statistics by simulation under rather general conditions. Given the observations $(y_t, \mathbf{z}_t), t = 1, ..., T$, the log-likelihood of (17) has the form

$$L_T(\psi) = c - (T/2)\ln\sigma^2 - (1/2\sigma^2)\sum_{t=1}^{T}\{y_t - \phi'\mathbf{z}_t - \theta'\mathbf{z}_t G(\gamma; \mathbf{s}_t)\}^2$$

$\psi = (\phi', \theta')'$. Assuming γ known, the average score for the parameters in the conditional mean equals

$$\mathbf{s}_T(\psi, \gamma) = (\sigma^2 T)^{-1}\sum_{t=1}^{T}(\mathbf{z}_t \otimes \begin{bmatrix} 1 & G(\gamma; \mathbf{s}_t) \end{bmatrix}')\varepsilon_t. \tag{20}$$

Lagrange multiplier and Wald tests can be defined using (20) in the usual way. The LM test statistic equals

$$S_T^{\text{LM}}(\gamma) = T\mathbf{s}_T(\widetilde{\psi}, \gamma)'\widetilde{\mathbf{I}}_T(\widetilde{\psi}, \gamma)^{-1}\mathbf{s}_T(\widetilde{\psi}, \gamma)$$

where $\widetilde{\psi}$ is the maximum likelihood estimator of ψ under H_0 and $\widetilde{\mathbf{I}}_T(\widetilde{\psi}, \gamma)$ is a consistent estimator of the population information matrix $\mathbf{I}(\psi, \gamma)$. An empirical distribution of $S_T^{\text{LM}}(\gamma)$ is obtained by simulation as follows:

1. Generate T observations $\varepsilon_t^{(j)}, t = 1, ..., T$ for each $j = 1, ..., J$ from a normal $(0,\widetilde{\sigma}^2)$ distribution, JT observations in all.

2. Compute $\mathbf{s}_T^{(j)}(\psi, \gamma_a) = T^{-1}\sum_{t=1}^{T}(\mathbf{z}_t \otimes \left[\begin{array}{cc} 1 & G(\gamma_a; \mathbf{s}_t) \end{array}\right]')u_t^{(j)}$ where $\gamma_a \in \Gamma_A \subset \Gamma$.

3. Set $S_T^{\mathrm{LM}(j)}(\gamma_a) = T\mathbf{s}_T^{(j)}(\widetilde{\psi}, \gamma_a)'\widetilde{\mathbf{I}}_T^{(j)}(\widetilde{\psi}, \gamma_a)^{-1}\mathbf{s}_T^{(j)}(\widetilde{\psi}, \gamma_a)$.

4. Compute $S_T^{\mathrm{LM}(j)}$ from $S_T^{\mathrm{LM}(j)}(\gamma_a), a = 1, ..., A$.

Carrying out these steps once gives a simulated value of the statistic. By repeating them J times one generates a random sample $\{S_T^{\mathrm{LM}(1)}, ..., S_T^{\mathrm{LM}(J)}\}$ from the null distribution of S_T^{LM}. If the value of S_T^{LM} obtained directly from the sample exceeds the $100(1-\alpha)\%$ quantile of the empirical distribution, the null hypothesis is rejected at (approximately) significance level α. The power of the test depends on the quality of the approximation Γ_A. Hansen (1996) applied this technique to testing linearity against the two-regime threshold autoregressive model. The empirical distribution may also be obtained by bootstrapping the residuals of the null model.

There is another way of handling the identification problem that is applicable in the context of STR models. Instead of approximating the unknown distribution of a test statistic it is possible to approximate the conditional log-likelihood or the nonlinear model in such a way that the identification problem is circumvented. See Luukkonen, Saikkonen and Teräsvirta (1988), Granger and Teräsvirta (1993) and Teräsvirta (1994) for discussion. Define $\gamma = (\gamma_1, \gamma_2')'$ in (17) and assume that $G(\gamma_1, \gamma_2; \mathbf{s}_t) \equiv 0$ for $\gamma_1 = 0$. Assume, furthermore, that $G(\gamma_1, \gamma_2; \mathbf{s}_t)$ is at least k times continuously differentiable for all values of $\mathbf{s}_t$ and γ.

It is now possible to approximate the transition function by a Taylor expansion and circumvent the identification problem. First note that due to lack of identification, the linearity hypothesis can also be expressed as $\mathrm{H}_0 : \gamma_1 = 0$. Function G is approximated locally around the null hypothesis as follows:

$$G(\gamma_1, \gamma_2; \mathbf{s}_t) = \sum_{j=1}^{k}(\gamma_1^j/j!)\delta_j(\mathbf{s}_t) + R_k(\gamma_1, \gamma_2; \mathbf{s}_t) \tag{21}$$

where $\delta_j(\mathbf{s}_t) = \frac{\partial^j}{\partial\gamma_1^j}G(\gamma_1, \gamma_2; \mathbf{s}_t)|_{\gamma_1=0}, j = 1, ..., k$. Replacing G in (17) by (21) yields, after reparameterization,

$$y_t = \phi'\mathbf{z}_t + \sum_{j=1}^{k}\theta_j(\gamma_1)'\mathbf{z}_t\delta_j(\mathbf{s}_t) + \varepsilon_t^* \tag{22}$$

where the parameter vectors $\theta_j(\gamma_1) = 0$ for $\gamma_1 = 0$, and the error term $\varepsilon_t^* = \varepsilon_t + \theta' \mathbf{z}_t R_k(\gamma_1, \gamma_2; \mathbf{s}_t)$. The original null hypothesis can now be restated as $\mathrm{H}_0' : \theta_j(\gamma_1) = 0, j = 1, ..., k$. It is a linear hypothesis in a linear model and can thus be tested using standard asymptotic theory, because under the null hypothesis $\varepsilon_t^* = \varepsilon_t$. Note, however, that this requires the existence of $\mathsf{E}\delta_j(\mathbf{s}_t)^2 \mathbf{z}_t \mathbf{z}_t'$. The auxiliary regression (22) can be viewed as a result of a trade-off in which information about the structural form of the alternative model is exchanged against a larger null hypothesis and standard asymptotic theory.

As an example, consider the STR model (3) and (4) and assume $K = 1$ in (4). It is a special case of (17) where $\gamma_2 = c$ and

$$G(\gamma_1, c; s_t) = (1 + \exp\{-\gamma_1(s_t - c)\})^{-1}, \gamma_1 > 0. \tag{23}$$

When $\gamma_1 = 0$, $G(\gamma_1, c; s_t) \equiv 1/2$. The first-order Taylor expansion of the transition function around $\gamma_1 = 0$ is

$$T(\gamma_1; s_t) = (1/2) - (\gamma_1/4)(s_t - c) + R_1(\gamma_1; s_t). \tag{24}$$

Substituting (24) for (23) in (17) yields, after reparameterization,

$$y_t = (\phi_0^*)' \mathbf{z}_t + (\phi_1^*)' \mathbf{z}_t s_t + \varepsilon_t^* \tag{25}$$

where $\phi_1^* = \gamma_1 \overline{\phi}_1^*$ such that $\overline{\phi}_1^* \neq \mathbf{0}$. The transformed null hypothesis is thus $\mathrm{H}_0' : \phi_1^* = \mathbf{0}$. Under this hypothesis and assuming that $\mathsf{E}s_t^2 \mathbf{z}_t \mathbf{z}_t'$ exists, the resulting LM statistic has an asymptotic χ^2 distribution with m degrees of freedom. This computationally simple test also has power against SR model, but Hansen's test that is designed directly against that alternative, is of course the more powerful of the two.

3.2 Building STR models

The STR model nests a linear regression model and is not identified when the data-generating process is the linear model. For this reason, a natural first step in building STR models is testing linearity against STR. There exists a data-based modelling strategy that consists of the three stages already mentioned: specification, estimation, and evaluation. It is described, among others, in Teräsvirta (1998), see also van Dijk, Teräsvirta and Franses (2002) or Teräsvirta (2004). Specification consists of testing linearity and, if rejected, determining the transition variable s_t. This is done using testing linearity against STR models with different transition variables. In the univariate case, determining the transition variable amounts to choosing the lag

y_{t-d}. The decision to select the type of the STR model (LSTR1 or LSTR2) is also made at the specification stage and is based on the results of a short sequence of tests within an auxiliary regression that is used for testing linearity; see Teräsvirta (1998) for details.

Specification is partly intertwined with estimation, because the model may be reduced by setting coefficients to zero according to some rule and re-estimating the reduced model. This implies that one begins with a large STR model and then continues 'from general to specific'. At the evaluation stage the estimated STR model is subjected to misspecification tests such as tests of no error autocorrelation, no autoregressive conditional heteroskedasticity, no remaining nonlinearity and parameter constancy. The tests are described in Teräsvirta (1998). A model that passes the in-sample tests can be used for out-of-sample forecasting.

The presence of unidentified nuisance parameters is also a problem in misspecification testing. The alternatives to the STR model in tests of no remaining nonlinearity and parameter constancy are not identified when the null hypothesis is valid. The identification problem is again circumvented using a Taylor series expansion. In fact, the linearity test applied at the specification stage can be viewed as a special case of the misspecification test of no remaining nonlinearity.

It may be mentioned that Medeiros, Teräsvirta and Rech (in press) constructed a similar strategy for modelling with neural networks. There the specification stage involves, except testing linearity, selecting the variables and the number of hidden units. Teräsvirta, Lin and Granger (1993) presented a linearity test against the neural network model using the Taylor series expansion idea; for a different approach, see Lee, White and Granger (1993).

In some forecasting experiments, STAR models have been fitted to data without first testing linearity, and assuming the structure of the model known in advance. As already discussed, this should lead to forecasts that are inferior to forecasts obtained from models that have been specified using data. The reason is that if the data-generating process is linear, the parameters of the STR or STAR model are not estimated consistently. This in turn must have a negative effect on forecasts, compared to models obtained by a specification strategy in which linearity is tested before attempting to build an STR or STAR model.

3.3 Building switching regression models

The switching regression model shares with the STR model the property that it nests a linear regression model and is not identified when the nested model

generates the observations. This suggests that a first step in specifying the switching regression model or the threshold autoregressive model should be testing linearity. In other words, one would begin by choosing between one and two regimes in (6). When this is done, it is usually assumed that the error variances in different regimes are the same: $\sigma_j^2 \equiv \sigma^2$, $j = 1, ..., r$.

More generally, the specification stage consists of selecting both the switching variable s_t and determining the number of regimes. There are several ways of determining the number of regimes. Hansen (1999) suggested a sequential testing approach to the problem. He discussed the SETAR model, but his considerations apply to the multivariate model as well. Hansen (1999) suggested a likelihood ratio test for this situation and showed how inference can be conducted using an empirical null distribution of the test statistic generated by the bootstrap. Applied sequentially and starting from a linear model, Hansen's empirical-distribution based likelihood ratio test can in principle be used for selecting the number of regimes in a SETAR model.

The test has excellent size and power properties as a linearity test, but it does not always work as well as a sequential test in the SETAR case. Suppose that the true model has three regimes, and Hansen's test is used for testing two regimes against three. Then it may happen that the estimated model with two regimes generates explosive realizations, although the data-generating process with three regimes is stationary. This causes problems in bootstrapping the test statistic under the null hypothesis. If the model is a static switching regression model, this problem does not occur.

Gonzalo and Pitarakis (2002) designed a technique based on model selection criteria. The number of regimes is chosen sequentially. Expanding the model by adding another regime is discontinued when the value of the model selection criterion, such as BIC, does not decrease any more. A drawback of this technique is that the significance level of each individual comparison (j regimes vs. $j + 1$) is a function of the size of the model and cannot be controlled by the model builder. This is due to the fact that the size of the penalty in the model selection criterion is a function of the number of parameters in the two models under comparison.

Recently, Strikholm and Teräsvirta (2005) suggested approximating the threshold autoregressive model by a multiple STAR model with a large fixed value for the slope parameter γ. The idea is then to first apply the linearity test and then the test of no remaining nonlinearity sequentially to find the number of regimes. This gives the modeller an approximate control over the significance level, and the technique appears to work reasonably well in simulations. Selecting the switching variable s_t can be incorporated into every one of these three approaches; see, for example, Hansen (1999).

Estimation of parameters is carried out by forming a grid of values for

the threshold parameter, estimating the remaining parameters conditionally on this value for each value in the grid and minimizing the sum of squared errors.

The likelihood ratio test of Hansen (1999) can be regarded as a misspecification test of the estimated model. The estimated model can also be tested following the suggestion by Eitrheim and Teräsvirta (1996) that is related to the ideas in Strikholm and Teräsvirta (2005). One can re-estimate the threshold autoregressive model as a STAR model with a large fixed γ and apply misspecification tests developed for the STAR model. Naturally, in this case there is no asymptotic distribution theory for these tests but they may nevertheless serve as useful indicators of misspecification. Tong (1990, Section 5.6) discusses ways of checking the adequacy of estimated nonlinear models that also apply to SETAR models.

3.4 Building Markov-switching regression models

The MS regression model has a structure similar to the previous models in the sense that it nests a linear model, and the model is not identified under linearity. In that case the transition probabilities are unidentified nuisance parameters. The first stage of building MS regression models should therefore be testing linearity. Nevertheless, this is very rarely the case in practice. An obvious reason is that testing linearity against the MS-AR alternative is computationally demanding. Applying the general theory of Hansen (1996) to this testing problem would require more computations than it does when the alternative is a threshold autoregressive model. Garcia (1998) offers an alternative that is computationally less demanding but does not appear to be in common use. Most practitioners fix the number of regimes in advance, and the most common choice appears to be two regimes. For an exception to this practice, see Li and Xu (2002).

Estimation of Markov-switching models is more complicated than estimation of models described in previous sections. This is because the model contains two unobservable processes: the Markov chain indicating the regime and the error process ε_t. Hamilton (1993) and Hamilton (1994, Chapter 22), among others, discussed maximum likelihood estimation of parameters in this framework.

Misspecification tests exist for the evaluation of Markov-switching models. The tests proposed in Hamilton (1996) are Lagrange multiplier tests. If the model is a regression model, a test may be constructed for testing whether there is autocorrelation or ARCH effects in the process or whether a higher-order Markov chain would be necessary to adequately characterize the dynamic behaviour of the switching process.

Breunig, Najarian and Pagan (2003) consider other types of tests and give examples of their use. These include consistency tests for finding out whether assumptions made in constructing the Markov-switching model are compatible with the data. Furthermore, they discuss encompassing tests that are used to check whether a parameter of some auxiliary model can be encompassed by the estimated Markov-switching model. The authors also emphasize the use of informal graphical methods in checking the validity of the specification. These methods can be applied to other nonlinear models as well.

4 Forecasting with nonlinear models

4.1 Analytical point forecasts

For some nonlinear models, forecasts for more than one period ahead can be obtained analytically. This is true for many nonlinear moving average models that are linear in parameters. As an example, consider the asymmetric moving average model (16) , assume that it is invertible, and set $q = 2$ for simplicity. The optimal point forecast one period ahead equals

$$y_{t+1|t} = \mathsf{E}\{y_{t+1}|\mathcal{F}_t\} = \mu + \theta_1\varepsilon_t + \theta_2\varepsilon_{t-1} + \psi_1 I(\varepsilon_t > 0)\varepsilon_t + \psi_2 I(\varepsilon_{t-1} > 0)\varepsilon_{t-1}$$

and two periods ahead

$$y_{t+2|t} = \mathsf{E}\{y_{t+2}|\mathcal{F}_t\} = \mu + \theta_2\varepsilon_t + \psi_1 \mathsf{E}I(\varepsilon_{t+1} > 0)\varepsilon_{t+1} + \psi_2 I(\varepsilon_t > 0)\varepsilon_t.$$

For example, if $\varepsilon_t \sim \text{nid}(0, \sigma^2)$, then $\mathsf{E}I(\varepsilon_t > 0)\varepsilon_t = (\sigma^2/2)\sqrt{\pi/2}$. For more than two periods ahead, the forecast is simply the unconditional mean of y_t :

$$\mathsf{E}y_t = \mu + (\psi_1 + \psi_2)\mathsf{E}I(\varepsilon_t > 0)\varepsilon_t$$

exactly as in the case of a linear MA(2) model.

Another nonlinear model from which forecasts can be obtained using analytical expressions is the Markov-switching model. Consider model (8) and suppose that the exogenous variables are generated by the following linear model:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \eta_{t+1}. \tag{26}$$

The conditional expectation of y_{t+1}, given the information up until t from

(8) , has the form

$$\begin{aligned}\mathsf{E}\{y_{t+1}|\mathbf{x}_t,\mathbf{w}_t\} &= \mathsf{E}[\sum_{j=1}^{r}\{y_{t+1}|\mathbf{x}_t,\mathbf{w}_t,s_{t+1}=j\}]\Pr\{s_{t+1}=j|\mathbf{x}_t,\mathbf{w}_t\}\\ &= \sum_{j=1}^{r} p_{j,t+1}(\alpha_{1j}'\mathbf{A}\mathbf{x}_t+\alpha_{2j}'\mathbf{w}_t)\end{aligned}$$

where $p_{j,t+1}$= $\Pr\{s_{t+1}=j|\mathbf{x}_t,\mathbf{w}_t\}$, is the conditional probability of the process being in state j at time $t+1$ given the past observable information. Then the forecast of y_{t+1} given $\mathbf{x}_t$ and $\mathbf{w}_t$ and involving the forecasts of $p_{j,t+1}$ becomes

$$y_{t+1|t} = \sum_{j=1}^{r} p_{j,t+1|t}(\alpha_{1j}'\mathbf{A}\mathbf{x}_t+\alpha_{2j}'\mathbf{w}_t). \tag{27}$$

In (27) , $p_{j,t+1|t} = \Pr\{s_{t+1}=j|\mathbf{x}_t,\mathbf{w}_t\}$ is a forecast of $p_{j,t+1}$ from $\mathbf{p}_{t+1|t}' = \mathbf{p}_t'\mathbf{P}$ where $\mathbf{p}_t = (p_{1,t},...,p_{r,t})'$ with $p_{j,t} = \Pr\{s_t = j|\mathbf{x}_t,\mathbf{w}_t\}$, $j = 1,...,r$, and $\mathbf{P}$ =$[p_{ij}]$ is the matrix of transition probabilities defined in (9).

Generally, the forecast for $h \geq 2$ steps ahead has the following form

$$y_{t+h|t} = \sum_{j=1}^{r} p_{j,t+h|t}(\alpha_{1j}'\mathbf{A}^h\mathbf{x}_t+\alpha_{2j}'\mathbf{w}_{t+h-1}^*)$$

where the forecasts $p_{j,t+h|t}$ of the regime probabilities are obtained from the relationship $\mathbf{p}_{t+h|t}' = \mathbf{p}_t'\mathbf{P}^h$ with $\mathbf{p}_{t+h|t} = (p_{1,t+h|t},...,p_{r,t+h|t})'$ and $\mathbf{w}_{t+h-1}^* = (y_{t+h-1|t},...,y_{t+1|t},y_t,...,y_{t-p+h-1})'$, $h \geq 2$.

As a simple example, consider the first-order autoregressive MS or SCAR model with two regimes

$$y_t = \sum_{j=1}^{2}(\phi_{0j}+\phi_{1j}y_{t-1})I(s_t=j)+\varepsilon_t \tag{28}$$

where $\varepsilon_t \sim \text{nid}(0,\sigma^2)$. From (28) it follows that the one-step-ahead forecast equals

$$y_{t+1|t} = \mathsf{E}\{y_{t+1}|y_t\} = \mathbf{p}_t'\mathbf{P}\phi_0+\mathbf{p}_t'\mathbf{P}\phi_1 y_t$$

where $\phi_j = (\phi_{j1},\phi_{j2})'$, $j = 0,1$. For two steps ahead, one obtains

$$\begin{aligned}y_{t+2|t} &= \mathbf{p}_t'\mathbf{P}^2\phi_0+\mathbf{p}_t'\mathbf{P}^2\phi_1 y_{t+1|t}\\ &= \mathbf{p}_t'\mathbf{P}^2\phi_0+(\mathbf{p}_t'\mathbf{P}^2\phi_1)(\mathbf{p}_t'\mathbf{P}\phi_0)+(\mathbf{p}_t'\mathbf{P}^2\phi_1)(\mathbf{p}_t'\mathbf{P}\phi_1)y_t.\end{aligned}$$

Generally, the h-step ahead forecast, $h \geq 2$, has the form

$$y_{t+h|t} = \mathbf{p}_t'\mathbf{P}^h\phi_0 + \sum_{i=0}^{h-2}(\prod_{j=0}^{i}\mathbf{p}_t'\mathbf{P}^{h-j}\phi_1)\mathbf{p}_t'\mathbf{P}^{h-i-1}\phi_0 + (\prod_{j=1}^{h}\mathbf{p}_t'\mathbf{P}^j\phi_1)y_t.$$

Thus all forecasts can be obtained analytically by a sequence of linear operations. This is a direct consequence of the fact that the regimes in (8) are linear in parameters. If they were not, the situation would be different. This would also be the case if the exogenous variables were generated by a nonlinear process instead of the linear model (26). Forecasting in such situations will be considered next.

4.2 Numerical techniques in forecasting

Forecasting for more than one period ahead with nonlinear models such as the STR or SR model requires numerical techniques. Granger and Teräsvirta (1993, Chapter 9), Lundbergh and Teräsvirta (2002), Franses and van Dijk (2000) and Fan and Yao (2003), among others, discuss ways of obtaining such forecasts. In the following discussion, it is assumed that the nonlinear model is correctly specified. In practice, this is not the case. Recursive forecasting that will be considered here may therefore lead to rather inaccurate forecasts if the model is badly misspecified. Evaluation of estimated models by misspecification tests and other means before forecasting with them is therefore important.

Consider the following simple nonlinear model

$$y_t = g(\mathbf{x}_{t-1};\theta) + \varepsilon_t \tag{29}$$

where $\varepsilon_t \sim \text{iid}(0,\sigma^2)$ and $\mathbf{x}_t$ is a $(k \times 1)$ vector of exogenous variables. Forecasting one period ahead does not pose any problem, for the forecast

$$y_{t+1|t} = \mathsf{E}(y_{t+1}|\mathbf{x}_t) = g(\mathbf{x}_t;\theta).$$

We bypass an extra complication by assuming that θ is known, which means that the uncertainty from the estimation of parameters is ignored. Forecasting two steps ahead is already a more complicated affair because we have to work out $\mathsf{E}(y_{t+2}|\mathbf{x}_t)$. Suppose we can forecast $\mathbf{x}_{t+1}$ from the linear first-order vector autoregressive model

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \eta_{t+1} \tag{30}$$

where $\eta_t = (\eta_{1t}, ..., \eta_{kt})' \sim \text{iid}(\mathbf{0}, \mathbf{\Sigma}_\eta)$. The one-step-ahead forecast of $\mathbf{x}_{t+1}$ is $\mathbf{x}_{t+1|t} = \mathbf{A}\mathbf{x}_t$. This yields

$$\begin{aligned} y_{t+2|t} &= \mathsf{E}(y_{t+2}|\mathbf{x}_t) = \mathsf{E}g(\mathbf{A}\mathbf{x}_t + \eta_{t+1}; \theta) \\ &= \int_{\eta_1} ... \int_{\eta_k} g(\mathbf{A}\mathbf{x}_t + \eta_{t+1}; \theta) \mathrm{d}F(\eta_1, ..., \eta_k) \end{aligned} \tag{31}$$

which is a k-fold integral and where $F(\eta_1, ..., \eta_k)$ is the joint cumulative distribution function of η_t. Even in the simple case where $\mathbf{x}_t = (y_t, ..., y_{t-p+1})'$ one has to integrate out the error term ε_t from the expected value $\mathsf{E}(y_{t+2}|\mathbf{x}_t)$. It is possible, however, to ignore the error term and just use

$$y^S_{t+2|t} = g(\mathbf{x}_{t+1|t}; \theta)$$

which Tong (1990) calls the 'skeleton' forecast. This method, while easy to apply, yields, however, a biased forecast for y_{t+2}. It may lead to substantial losses of efficiency; see Lin and Granger (1994) for simulation evidence of this.

On the other hand, numerical integration of (31) is tedious. Granger and Teräsvirta (1993) call this method of obtaining the forecast the exact method, as opposed to two numerical techniques that can be used to approximate the integral in (31). One of them is based on simulation, the other one on bootstrapping the residuals $\{\widehat{\eta}_t\}$ of the estimated equation (30) or the residuals $\{\widehat{\varepsilon}_t\}$ of the estimated model (29) in the univariate case. In the latter case the parameter estimates thus do have a role to play, but the additional uncertainty of the forecasts arising from the estimation of the model is not accounted for.

The simulation approach requires that a distributional assumption is made about the errors η_t. One draws a sample of N independent error vectors $\{\eta^{(1)}_{t+1}, ..., \eta^{(N)}_{t+1}\}$ from this distribution and computes the Monte Carlo forecast

$$y^{MC}_{t+2|t} = (1/N) \sum_{i=1}^{N} g(\mathbf{x}_{t+1|t} + \eta^{(i)}_{t+1}; \theta). \tag{32}$$

The bootstrap forecast is similar to (32) and has the form

$$y^B_{t+2|t} = (1/N_B) \sum_{i=1}^{N_B} g(\mathbf{x}_{t+1|t} + \widehat{\eta}^{(i)}_{t+1}; \theta) \tag{33}$$

where the errors $\{\widehat{\eta}^{(1)}_{t+1}, ..., \widehat{\eta}^{(N_B)}_{t+1}\}$ have been obtained by drawing them from the set of estimated residuals of model (30) with replacement. The difference

between (32) and (33) is that the former is based on an assumption about the distribution of η_{t+1}, whereas the latter does not make use of a distributional assumption. It requires, however, that the error vectors are assumed independent.

This generalizes to longer forecast horizons: For example,

$$\begin{aligned} y_{t+3|t} &= \mathsf{E}(y_{t+3}|\mathbf{x}_t) = \mathsf{E}\{g(\mathbf{x}_{t+2};\theta)|\mathbf{x}_t\} \\ &= \mathsf{E}\{g(\mathbf{A}\mathbf{x}_{t+1}+\eta_{t+2};\theta)|\mathbf{x}_t\} = \mathsf{E}g(\mathbf{A}^2\mathbf{x}_t+\mathbf{A}\eta_{t+1}+\eta_{t+2};\theta) \\ &= \int_{\eta_1^{(2)}} ... \int_{\eta_k^{(2)}} \int_{\eta_1^{(1)}} ... \int_{\eta_k^{(1)}} g(\mathbf{A}^2\mathbf{x}_t+\mathbf{A}\eta_{t+1}+\eta_{t+2};\theta) \\ &\times \mathrm{d}F(\eta_1^{(1)},...,\eta_k^{(1)},\eta_1^{(2)},...,\eta_k^{(2)}) \end{aligned}$$

which is a $2k$-fold integral. Calculation of this expectation by numerical integration may be a huge task, but simulation and bootstrap approaches are applicable. In the general case where one forecasts h steps ahead and wants to obtain the forecasts by simulation, one generates the random variables $\eta_{t+1}^{(i)},...,\eta_{t+h}^{(i)}$, $i=1,...,N$, and sequentially computes N forecasts for $y_{t+1|t},...,y_{t+h|t}$, $h \geq 2$. These are combined to a single point forecast for each of the time-points by simple averaging as in (32). Bootstrap-based forecasts can be computed in an analogous fashion.

If the model is univariate, the principles do not change. Consider, for simplicity, the following stable first-order autoregressive model

$$y_t = g(y_{t-1};\theta) + \varepsilon_t \tag{34}$$

where $\{\varepsilon_t\}$ is a sequence of independent, identically distributed errors such that $\mathsf{E}\varepsilon_t = 0$ and $\mathsf{E}\varepsilon_t^2 = \sigma^2$. In that case,

$$\begin{aligned} y_{t+2|t} &= \mathsf{E}[g(y_{t+1};\theta)+\varepsilon_{t+2}|y_t] = \mathsf{E}g(g(y_t;\theta)+\varepsilon_{t+1};\theta) \\ &= \int_{\varepsilon} g(g(y_t;\theta)+\varepsilon);\theta)\mathrm{d}F(\varepsilon) \end{aligned} \tag{35}$$

The only important difference between (31) and (35) is that in the latter case, the error term that has to be integrated out is the error term of the autoregressive model (34). In the former case, the corresponding error term is the error term of the vector process (30) , and the error term of (29) need not be simulated. For an example of a univariate case, see Lundbergh and Teräsvirta (2002).

It should be mentioned that there is an old strand of literature on forecasting from nonlinear static simultaneous-equation models in which the techniques just presented are discussed and applied. The structural equations of the model have the form

$$\mathbf{f}(\mathbf{y}_t,\mathbf{x}_t,\theta) = \varepsilon_t \tag{36}$$

where $\mathbf{f}$ is an $n \times 1$ vector of functions of the n endogenous variables $\mathbf{y}_t$, $\mathbf{x}_t$ is a vector of exogenous variables, $\{\varepsilon_t\}$ a sequence of independent error vectors, and θ the vector of parameters. It is assumed that (36) implicitly defines a unique inverse relationship

$$\mathbf{y}_t = \mathbf{g}(\varepsilon_t, \mathbf{x}_t, \theta).$$

There may not exist a closed form for $\mathbf{g}$ or the conditional mean and covariance matrix of $\mathbf{y}_t$. Given $\mathbf{x}_t = \mathbf{x}^0$, the task is to forecast $\mathbf{y}_t$. Different assumptions on ε_t lead to skeleton or "deterministic" forecasts, exact or "closed form" forecasts, or Monte Carlo forecasts; see Brown and Mariano (1984). The order of bias in these forecasts has been a topic of discussion, and Brown and Mariano showed that the order of bias in skeleton forecasts is O(1).

4.3 Forecasting using recursion formulas

It is also possible to compute forecasts numerically applying the Chapman-Kolmogorov equation that can be used for obtaining forecasts recursively by numerical integration. Consider the following stationary first-order nonlinear autoregressive model

$$y_t = k(y_{t-1}; \theta) + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a sequence of iid$(0,\sigma^2)$ variables and that the conditional densities of the y_t are well-defined. Then a special case of the Chapman-Kolmogorov equation has the form, see for example Tong (1990, p. 346) or Franses and van Dijk (2000, p. 119-120)

$$f(y_{t+h}|y_t) = \int_{-\infty}^{\infty} f(y_{t+h}|y_{t+1}) f(y_{t+1}|y_t) \mathrm{d}y_{t+1}. \tag{37}$$

From (37) it follows that

$$y_{t+h|t} = \mathsf{E}\{y_{t+h}|y_t\} = \int_{-\infty}^{\infty} \mathsf{E}\{y_{t+h}|y_{t+1}\} f(y_{t+1}|y_t) \mathrm{d}y_{t+1} \tag{38}$$

which shows how $\mathsf{E}\{y_{t+h}|y_t\}$ may be obtained recursively. Consider the case $h = 2$. It should be noted that in (38), $f(y_{t+1}|y_t) = g(y_{t+1} - k(y_t;\theta)) = g(\varepsilon_{t+1})$. In order to calculate $f(y_{t+h}|y_t)$, one has to make an appropriate assumption about the error distribution $g(\varepsilon_{t+1})$. Since $\mathsf{E}\{y_{t+2}|y_{t+1}\} = k(y_{t+1};\theta)$, the forecast

$$y_{t+2|t} = \mathsf{E}\{y_{t+2}|y_t\} = \int_{-\infty}^{\infty} k(y_{t+1};\theta) g(y_{t+1} - k(y_t;\theta)) \mathrm{d}y_{t+1} \tag{39}$$

is obtained from (39) by numerical integration. For $h > 2$, one has to make use of both (38) and (39) . First, write

$$\mathsf{E}\{y_{t+3}|y_t\} = \int_{-\infty}^{\infty} k(y_{t+2};\theta) f(y_{t+2}|y_t) \mathrm{d}y_{t+2} \tag{40}$$

then obtain $f(y_{t+2}|y_t)$ from (37) where $h = 2$ and

$$f(y_{t+2}|y_{t+1}) = g(y_{t+2} - k(y_{t+1};\theta)).$$

Finally, the forecast is obtained from (40) by numerical integration.

It is seen that this method is computationally demanding for large values of h. Simplifications to alleviate the computational burden exist, see De Gooijer and De Bruin (1998). The latter authors consider forecasting with SETAR models with the normal forecasting error (NFE) method. As an example, take the first-order SETAR model

$$y_t = (\alpha_{01} + \alpha_{11} y_{t-1} + \varepsilon_{1t}) I(y_{t-1} < c) + (\alpha_{02} + \alpha_{12} y_{t-1} + \varepsilon_{2t}) I(y_{t-1} \geq c) \tag{41}$$

where $\{\varepsilon_{jt}\} \sim \mathrm{nid}(0, \sigma_j^2)$, $j = 1, 2$. For the SETAR model (41), the one-step-ahead minimum mean-square error forecast has the form

$$y_{t+1|t} = \mathsf{E}\{y_{t+1}|y_t < c\} I(y_t < c) + \mathsf{E}\{y_{t+1}|y_t \geq c\} I(y_t \geq c)$$

where $\mathsf{E}\{y_{t+1}|y_t < c\} = \alpha_{01} + \alpha_{11} y_t$ and $\mathsf{E}\{y_{t+1}|y_t \geq c\} = \alpha_{02} + \alpha_{12} y_t$. The corresponding forecast variance

$$\sigma_{t+1|t}^2 = \sigma_1^2 I(y_t < c) + \sigma_2^2 I(y_t \geq c).$$

From (41) it follows that the distribution of y_{t+1} given y_t is normal with mean $y_{t+1|t}$ and variance $\sigma_{t+1|t}^2$. Accordingly for $h \geq 2$, the conditional distribution of y_{t+h} given y_{t+h-1} is normal with mean $\alpha_{01} + \alpha_{11} y_{t+h-1}$ and variance σ_1^2 for $y_{t+h-1} < c$, and mean $\alpha_{02} + \alpha_{12} y_{t+h-1}$ and variance σ_2^2 for $y_{t+h-1} \geq c$. Let $z_{t+h-1|t} = (c - y_{t+h-1|t})/\sigma_{t+h-1|t}$ where $\sigma_{t+h-1|t}^2$ is the variance predicted for time $t + h - 1$. De Gooijer and De Bruin (1998) show that the h-steps ahead forecast can be approximated by the following recursive formula

$$\begin{aligned} y_{t+h|t} = {} & (\alpha_{01} + \alpha_{11} y_{t+h-1|t}) \Phi(z_{t+h-1|t}) + (\alpha_{02} + \alpha_{12} y_{t+h-1|t}) \Phi(-z_{t+h-1|t}) \\ & - (\alpha_{11} - \alpha_{21}) \sigma_{t+h-1|t} \phi(z_{t+h-1|t}) \end{aligned} \tag{42}$$

where $\Phi(x)$ is the cumulative distribution function of a standard normal variable x and $\phi(x)$ is the density function of x. The recursive formula for forecasting the variance is not reproduced here. The first two terms weight

the regimes together: the weights are equal for $y_{t+h-1|t} = c$. The third term is a "correction term" that depends on the persistence of the regimes and the error variances. This technique can be generalized to higher-order SETAR models. De Gooijer and De Bruin (1998) report that the NFE method performs well when compared to the exact method described above, at least in the case where the error variances are relatively small. They recommend the method as being very quick and easy to apply.

It may be expected, however, that the use of the methods described in this subsection will lose in popularity when increased computational power makes the simulation-based approach both quick and cheap to use.

4.4 Accounting for estimation uncertainty

In Sections 4.1 and 4.2 it is assumed that the parameters are known. In practice, the unknown parameters are replaced by their estimates and recursive forecasts are obtained using these estimates. There are two ways of accounting for parameter uncertainty. It may be assumed that the (quasi) maximum likelihood estimator $\widehat{\theta}$ of the parameter vector θ has an asymptotic normal distribution, that is,

$$\sqrt{T}(\widehat{\theta} - \theta) \xrightarrow{D} \mathsf{N}(\mathbf{0}, \mathbf{\Sigma}).$$

One then draws a new estimate from the $\mathsf{N}(\widehat{\theta}, T^{-1}\widehat{\Sigma})$ distribution and repeats the forecasting exercise with them. For recursive forecasting in Section 4.2 this means repeating the calculations in (32) M times. Confidence intervals for forecasts can then be calculated from the MN individual forecasts. Another possibility is to re-estimate the parameters using data generated from the original estimated model by bootstrapping the residuals, call the estimated model $\mathcal{M}_B$. The residuals of $\mathcal{M}_B$ are then used to recalculate (33), and this procedure is repeated M times. This is a computationally intensive procedure and, besides, because the estimated models have to be evaluated (for example, explosive ones have to be discarded, so they do not distort the results), the total effort is substantial. When the forecasts are obtained analytically as in Section 4.1, the computational burden is less heavy because the replications to generate (32) or (33) are avoided.

4.5 Interval and density forecasts

Interval and density forecasts are obtained as a by-product of computing forecasts numerically. The replications form an empirical distribution that can be appropriately smoothed to give a smooth forecast density. For surveys,

see Corradi and Swanson (2006) and Tay and Wallis (2002). As already mentioned, forecast densities obtained from nonlinear economic models may be asymmetric, which policy makers may find interesting. For example, if a density forecast of inflation is asymmetric suggesting that the error of the point forecast is more likely to be positive than negative, this may cause a policy response different from the opposite situation where the error is more likely to be negative than positive. The density may even be bi- or multimodal, although this may not be very likely in macroeconomic time series. For an example, see Lundbergh and Teräsvirta (2002), where the density forecast for the Australian unemployment rate four quarters ahead from an estimated STAR model, reported in Skalin and Teräsvirta (2002), shows some bimodality.

Density forecasts may be conveniently presented using fan charts; see Wallis (1999) and Lundbergh and Teräsvirta (2002) for examples. There are two ways of constructing fan charts. One, applied in Wallis (1999), is to base them on interquantile ranges. The other is to use highest density regions, see Hyndman (1996). The choice between these two depends on the forecaster's loss function. Note, however, that bi- or multimodal density forecasts are only visible in fan charts based on highest density regions.

Typically, the interval and density forecasts do not account for the estimation uncertainty, but see Corradi and Swanson (2006). Extending the considerations to do that when forecasting with nonlinear models would often be computationally very demanding. The reason is that estimating parameters of nonlinear models requires care (starting-values, convergence, etc.), and therefore simulations or bootstrapping involved could in many cases demand a large amount of both computational and human resources.

4.6 Combining forecasts

Forecast combination is a relevant topic in linear as well as in nonlinear forecasting. Combining nonlinear forecasts with forecasts from a linear model may sometimes lead to series of forecasts that are more robust (contain fewer extreme predictions) than forecasts from the nonlinear model. Following Granger and Bates (1969), the composite point forecast from models M_1 and M_2 is given by

$$\widehat{y}^{(1,2)}_{t+h|t} = (1-\lambda_t)\widehat{y}^{(1)}_{t+h|t} + \lambda_t \widehat{y}^{(2)}_{t+h|t} \tag{43}$$

where λ_t, $0 \leq \lambda_t \leq 1$, is the weight of the h-periods-ahead forecast $\widehat{y}^{(j)}_{t+h|t}$ of y_{t+h}. Suppose that the multi-period forecasts from these models are obtained numerically following the technique presented in Section 4.2. The same random numbers can be used to generate both forecasts, and combining the

forecasts simply amounts to combining each realization from the two models. This means that each one of the N pairs of simulated forecasts from the two models is weighted into a single forecast using weights λ_t (model M_2) and $1 - \lambda_t$ (model M_1). The empirical distribution of the N weighted forecasts is the combined density forecast from which one easily obtains the corresponding point forecast by averaging as discussed in Section 4.2.

Note that the weighting schemes themselves may be nonlinear functions of the past performance. This form of nonlinearity in forecasting is not discussed here, but see Deutsch, Granger and Teräsvirta (1994) for an application. The K-mean clustering approach to combining forecasts in Aiolfi and Timmermann (in press) is another example of a nonlinear weighting scheme. A detailed discussion of forecast combination and weighting schemes proposed in the literature can be found in Timmermann (2006).

4.7 Different models for different forecast horizons?

Multistep forecasting was discussed in Section 4.2 where it was argued that for most nonlinear models, multi-period forecasts have to be obtained numerically. While this is not nowadays computationally demanding, there may be other reasons for opting for analytically generated forecasts. They become obvious if one gives up the idea that the model assumed to generate the observations is the data-generating process. As already mentioned, if the model is misspecified, the forecasts from such a model are not likely to have any optimality properties, and another misspecified model may do a better job. The situation is illuminated by an example from Bhansali (2002). Suppose that at time T we want to forecast y_{T+2} from

$$y_t = \alpha y_{t-1} + \varepsilon_t \tag{44}$$

where $\mathsf{E}\varepsilon_t = 0$ and $\mathsf{E}\varepsilon_t\varepsilon_{t-j} = 0, j \neq 0$. Furthermore, y_T is assumed known. Then $y_{T+1|T} = \alpha y_T$ and $y_{T+2|T} = \alpha^2 y_T$, where $\alpha^2 y_T$ is the minimum mean square error forecast of y_{T+2} under the condition that (44) be the data-generating process. If this condition is not valid, the situation changes. It is also possible to forecast y_{T+2} directly from the model estimated by regressing y_t on y_{t-2}, the (theoretical) outcome being $y^*_{T+2|T} = \rho_2 y_T$ where $\rho_2 = \mathrm{corr}(y_t, y_{t-2})$. When model (44) is misspecified, $y^*_{T+2|T}$ obtained by the direct method may be preferred to $y_{T+2|T}$ in a linear least squares sense. The mean square errors of these two forecasts are equal if and only if $\alpha^2 = \rho_2$, that is, when the data-generating process is a linear AR(1)-process.

When this idea is applied to nonlinear models, the direct method has the advantage that no numerical generation of forecasts is necessary. The

forecasts can be produced exactly as in the one-step-ahead case. A disadvantage is that a separate model has to be specified and estimated for each forecast horizon. Besides, these models are also misspecifications of the data-generating process. In their extensive studies of forecasting macroeconomic series with linear and nonlinear models, Stock and Watson (1999) and Marcellino (2002) have used this method. The interval and density forecasts obtained this way may sometimes differ from the ones generated recursively as discussed in Section 4.2. In forecasting more than one period ahead, the recursive techniques allow asymmetric forecast densities. On the other hand, if the error distribution of the 'direct forecast' model is assumed symmetric around zero, density forecasts from such a model will also be symmetric densities.

Which one of the two approaches produces more accurate point forecasts is an empirical matter. Lin and Granger (1994) study this question by simulation. Two nonlinear models, the first-order STAR and the sign model, are used to generate the data. The forecasts are generated in three ways. First, they are obtained from the estimated model assuming that the specification was known. Second, a neural network model is fitted to the generated series and the forecasts produced with it. Third, the forecasts are generated from a nonparametric model fitted to the series. The focus is on forecasting two periods ahead. On the one hand, the forecast accuracy measured by the mean square forecast error deteriorates compared to the iterative methods (32) and (33) when the forecasts two periods ahead are obtained from a 'direct' STAR or sign model, i.e., from a model in which the first lag is replaced by a second lag. On the other hand, the direct method works much better when the model used to produce the forecasts is a neural network or a nonparametric model.

A recent large-scale empirical study by Marcellino, Stock and Watson (2004) addresses the question of choosing an appropriate approach in a linear framework, using 171 monthly US macroeconomic time series and forecast horizons up to 24 months. The conclusion is that obtaining the multi-step forecasts from a single model is preferable to the use of direct models. This is true in particular for longer forecast horizons. A comparable study involving nonlinear time series models does not as yet seem to be available.

5 Forecast accuracy

5.1 Comparing point forecasts

A frequently-asked question in forecasting with nonlinear models has been whether they perform better than linear models. While many economic phenomena and models are nonlinear, they may be satisfactorily approximated by a linear model, and this makes the question relevant. A number of criteria, such as the root mean square forecast error (RMSFE) or mean absolute error (MAE), have been applied for the purpose. It is also possible to test the null hypothesis that the forecasting performance of two models, measured in RMSFE or MAE or some other forecast error based criterion, is equally good against a one-sided alternative. This can be done for example by applying the Diebold-Mariano (DM) test; see Diebold and Mariano (1995) and Harvey, Leybourne and Newbold (1997). The test is not available, however, when one of the models nests the other. The reason is that when the data are generated from the smaller model, the forecasts are identical when the parameters are known. In this case the asymptotic distribution theory for the DM statistic no longer holds.

This problem is present in comparing linear and many nonlinear models, such as the STAR, SETAR or MS (SCAR) model, albeit in a different form. These models nest a linear model, but the nesting model is not identified when the smaller model has generated the observations. Thus, if the parameter uncertainty is accounted for, the asymptotic distribution of the DM statistic may depend on unknown nuisance parameters, and the standard distribution theory does not apply.

Solutions to the problem of nested models are discussed in detail in West (2006), and here the attention is merely drawn to two approaches. Recently, Corradi and Swanson (2002, 2004) have considered what they call a generic test of predictive accuracy. The forecasting performance of two models, a linear model (M_0) nested in a nonlinear model and the nonlinear model (M_1), is under test. Following Corradi and Swanson (2004), define the models as follows:

$$\mathsf{M}_0: \qquad y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_{0t}$$

where $(\phi_0, \phi_1)' = \arg\min_{(\phi_0,\phi_1)\in\mathbf{\Phi}} \mathsf{E}g(y_t - \phi_0 - \phi_1 y_{t-1})$. The alternative has the form

$$\mathsf{M}_1: \qquad y_t = \phi_0(\gamma) + \phi_1(\gamma) y_{t-1} + \phi_2(\gamma) G(\mathbf{w}_t; \gamma) + \varepsilon_{1t} \tag{45}$$

where, setting $\phi(\gamma) = (\phi_0(\gamma), \phi_1(\gamma), \phi_2(\gamma))'$,

$$\phi(\gamma) = \mathrm{argmin}_{\phi(\gamma)\in\mathbf{\Phi}(\gamma)} \mathsf{E}g(y_t - \phi_0(\gamma) - \phi_1(\gamma) y_{t-1} - \phi_2(\gamma) G(\mathbf{w}_t; \gamma))$$

Furthermore, $\gamma \in \Gamma$ is a $d \times 1$ vector of nuisance parameters and Γ a compact subset of $\mathcal{R}^d$. The loss function is the same as the one used in the forecast comparison: for example the mean square error. The logistic function (4) may serve as an example of the nonlinear function $G(\mathbf{w}_t; \gamma)$ in (45) .

The null hypothesis equals $\mathrm{H}_0 : \mathsf{E}g(\varepsilon_{0,t+1}) = \mathsf{E}g(\varepsilon_{1,t+1})$, and the alternative is $\mathrm{H}_1 : \mathsf{E}g(\varepsilon_{0,t+1}) > \mathsf{E}g(\varepsilon_{1,t+1})$. The null hypothesis corresponds to equal forecasting accuracy, which is achieved if $\phi_2(\gamma) = 0$ for all $\gamma \in \Gamma$. This allows restating the hypotheses as follows:

$$\begin{aligned} &H_0 : \phi_2(\gamma) = 0 \text{ for all } \gamma \in \Gamma \\ &H_1 : \phi_2(\gamma) \neq 0 \text{ for at least one } \gamma \in \Gamma. \end{aligned} \tag{46}$$

Under this null hypothesis,

$$\mathsf{E}g'(\varepsilon_{0,t+1})G(\mathbf{w}_t; \gamma) = 0 \text{ for all } \gamma \in \Gamma \tag{47}$$

where

$$g'(\varepsilon_{0,t}) = \frac{\partial g}{\partial \varepsilon_{0,t}} \frac{\partial \varepsilon_{0,t}}{\partial \phi} = -\frac{\partial g}{\partial \varepsilon_{0,t}}(1, y_{t-1}, G(\mathbf{w}_{t-1}; \gamma))'.$$

For example,if $g(\varepsilon) = \varepsilon^2$, then $\partial g/\partial \varepsilon = 2\varepsilon$. The values of $G(\mathbf{w}_t; \gamma)$ are obtained using a sufficiently fine grid. Now, equation (47) suggests a conditional moment test of type Bierens (1990) for testing (46). Let

$$\widehat{\phi}_T = (\widehat{\phi}_0, \widehat{\phi}_1)' = \arg\min_{\phi \in \Phi} T^{-1} \sum_{t=1}^{T} g(y_t - \phi_0 - \phi_1 y_{t-1})$$

and define $\widehat{\varepsilon}_{0,t+1|t} = y_{t+1} - \widehat{\phi}_t' \mathbf{y}_t$ where $\mathbf{y}_t = (1, y_t)'$, for $t = T, T+1, ..., T-1$. The test statistic is

$$M_P = \int_\Gamma m_P(\gamma)^2 w(\gamma) \mathrm{d}\gamma \tag{48}$$

where

$$m_P(\gamma) = T^{-1/2} \sum_{t=T}^{T+P-1} g'(\widehat{\varepsilon}_{0,t+1|t}) G(\mathbf{z}_t; \gamma)$$

and the absolutely continuous weight function $w(\gamma) \geq 0$ with $\int_\Gamma w(\gamma)\mathrm{d}\gamma = 1$. The (nonstandard) asymptotic distribution theory for M_P is discussed in Corradi and Swanson (2002).

Statistic (48) does not answer the same question as the DM statistic. The latter can be used for investigating whether a given nonlinear model yields more accurate forecasts than a linear model not nested in it. The former answers a different question: "Does a given *family* of nonlinear models have

a property such that one-step-ahead forecasts from models belonging to this family are more accurate than the corresponding forecasts from a linear model nested in it?"

Some forecasters who apply nonlinear models that nest a linear model begin by testing linearity against their nonlinear model. This practice is often encouraged; see, for example, Teräsvirta (1998). If one rejects the linearity hypothesis, then one should also reject (46) , and an out-of-sample test would thus appear redundant. In practice it is possible, however, that (46) is not rejected although linearity is. This may be the case if the nonlinear model is misspecified, or there is a structural break or smooth parameter change in the prediction period, or this period is so short that the test is not sufficiently powerful. The role of out-of-sample tests in forecast evaluation compared to in-sample tests has been discussed in Inoue and Kilian (2004).

If one wants to consider the original question which the Diebold-Mariano test was designed to answer, a new test, recently developed by Giacomini and White (2003), is available. This is a test of conditional forecasting ability as opposed to most other tests including the Diebold-Mariano statistic that are tests of unconditional forecasting ability. The test is constructed under the assumption that the forecasts are obtained using a moving data window: the number of observations in the sample used for estimation does not increase over time. It is operational under rather mild conditions that allow heteroskedasticity. Suppose that there are two models M_1 and M_2 such that

$$\mathsf{M}_j: \quad y_t = f^{(j)}(\mathbf{w}_t; \theta_j) + \varepsilon_{jt}, j = 1, 2$$

where $\{\varepsilon_{jt}\}$ is a martingale difference sequence with respect to the information set $\mathcal{F}_{t-1}$. The null hypothesis is

$$\mathsf{E}[\{g_{t+\tau}(y_{t+\tau}, \widehat{f}^{(1)}_{mt}) - g_{t+\tau}(y_{t+\tau}, \widehat{f}^{(2)}_{mt})\}|\mathcal{F}_{t-1}] = 0 \tag{49}$$

where $g_{t+\tau}(y_{t+\tau}, \widehat{f}^{(j)}_{mt})$ is the loss function, $\widehat{f}^{(j)}_{mt}$ is the τ-periods-ahead forecast for $y_{t+\tau}$ from model j estimated from the observations $t-m+1, ..., t$. Assume now that there exist T observations, $t = 1, ..., T$, and that forecasting is begun at $t = t_0 > m$. Then there will be $T_0 = T - \tau - t_0$ forecasts available for testing the null hypothesis.

Carrying out the test requires a test function $\mathbf{h}_t$ which is a $p \times 1$ vector. Under the null hypothesis, owing to the martingale difference property of the loss function difference,

$$\mathsf{E}\mathbf{h}_t \Delta g_{t+\tau} = \mathbf{0}$$

for all $\mathcal{F}$-measurable $p \times 1$ vectors $\mathbf{h}_t$. Bierens (1990) used a similar idea ($\Delta g_{t+\tau}$ replaced by a function of the error term ε_t) to construct a general

model misspecification test. The choice of test function $\mathbf{h}_t$ is left to the user, and the power of the test depends on it. Assume now that $\tau = 1$. The GW test statistic has the form

$$S_{T_0,m} = T_0(T_0^{-1}\sum_{t=t_0}^{T_0}\mathbf{h}_t\Delta g_{t+\tau})'\widehat{\mathbf{\Omega}}_{T_0}^{-1}(T_0^{-1}\sum_{t=t_0}^{T_0}\mathbf{h}_t\Delta g_{t+\tau}) \tag{50}$$

where $\widehat{\mathbf{\Omega}}_{T_0} = T_0^{-1}\sum_{t=t_0}^{T_0}(\Delta g_{t+\tau})^2\mathbf{h}_t\mathbf{h}_t'$ is a consistent estimator of the covariance matrix $\mathsf{E}(\Delta g_{t+\tau})^2\mathbf{h}_t\mathbf{h}_t'$. When $\tau > 1$, $\widehat{\mathbf{\Omega}}_{T_0}$ has to be modified to account for correlation in the forecast errors; see Giacomini and White (2003). Under the null hypothesis (49) , the GW statistic (50) has an asymptotic χ^2-distribution with p degrees of freedom.

The GW test has not yet been applied to comparing the forecast ability of a linear model and a nonlinear model nested in it. Two things are important in applications. First, the estimation is based on a rolling window, but the size of the window may vary over time. Second, the outcome of the test depends on the choice of the test function $\mathbf{h}_t$. Elements of $\mathbf{h}_t$ not correlated with $\Delta g_{t+\tau}$ have a negative effect on the power of the test.

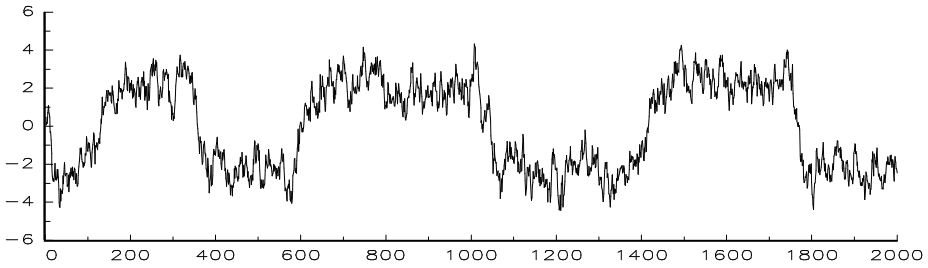
An important advantage with the GW test is that it can be applied to comparing methods for forecasting and not only models. The asymptotic distribution theory covers the situation where the specification of the model or models changes over time, which has sometimes been the case in practice. Swanson and White (1995,1997a,b) allow the specification to switch between a linear and a neural network model. In Teräsvirta et al. (2005), switches between linear on the one hand and nonlinear specifications such as the AR-NN and STAR model on the other are an essential part of their forecasting exercise.

6 Lessons from a simulation study

Building nonlinear time series models is generally more difficult than constructing linear models. A main reason for building nonlinear models for forecasting must therefore be that they are expected to forecast better than linear models. It is not certain, however, that this is so. Many studies, some of which will be discussed later, indicate that in forecasting macroeconomic series, nonlinear models may not forecast better than linear ones. In this section we point out that sometimes this may be the case even when the nonlinear model is the data-generating process.

As an example, we briefly review a simulation study in Lundbergh and Teräsvirta (2002). The authors generate 10^6 observations from the following

Figure 1 A realization of 2000 observations from model (51)



LSTAR model

$$y_t = -0.19 + 0.38(1 + \exp\{-10y_{t-1}\})^{-1} + 0.9y_{t-1} + 0.4\varepsilon_t \qquad (51)$$

where $\{\varepsilon_t\} \sim \text{nid}(0, 1)$. Model (51) may also be viewed as a special case of the neural network model (11) with a linear unit and a single hidden unit. The model has the property that the realization of 10^6 observations tends to fluctuate long periods around a local mean, either around -1.9 or 1.9. Occasionally, but not often, it switches from one 'regime' to the other, and the switches are relatively rapid. This is seen from Figure 1 that contains a realization of 2000 observations from (51) .

The authors fit the model with the same parameters as in (51) to a large number of subseries of 1000 observations, estimate the parameters, and forecast recursively up to 20 periods ahead. The results are compared to forecasts obtained from first-order linear autoregressive models fitted to the same subseries. The measure of accuracy is the relative efficiency (RE) measure of Mincer and Zarnowitz (1969), that is, the ratio of the RMSFEs of the two forecasts. It turns out that the forecasts from the LSTAR model are more efficient than the ones from the linear model: the RE measure moves from about 0.96 (one period ahead forecasts) to about 0.85 (20 periods ahead). The forecasts are also obtained assuming that the parameters are known: in that case the RE measure lies below 0.8 (20 periods ahead), so having to estimate the parameters affects the forecast accuracy as may be expected.

This is in fact not surprising, because the data-generating process is an LSTAR model. The authors were also interested in knowing how well this model forecasts when there is a large change in the value of the realization. This is defined as a change of at least equal to 0.2 in the absolute value of the transition function of (51) . It is a rare occasion and occurs only in about

0.6% of the observations. The question was posed, because Montgomery, Zarnowitz, Tsay and Tiao (1998) had shown that the nonlinear models of the US unemployment rate they considered performed better than the linear AR model when the unemployment increased rapidly but not elsewhere. Thus it was deemed interesting to study the occurrence of this phenomenon by simulation.

The results showed that the LSTAR model was better than the AR(1) model. The authors, however, also applied another benchmark, the first-order AR model for the differenced series, the ARI(1,1) model. This model was chosen as a benchmark because in the subseries of 1000 observations ending when a large change was observed, the unit root hypothesis, when tested using the augmented Dickey-Fuller test, was rarely rejected. A look at Figure 1 helps one understand why this is the case. Against the ARI(1,1) benchmark, the RE of the estimated LSTAR model was 0.95 at best, when forecasting three periods ahead, but RE exceeded unity for forecast horizons longer than 13 periods. There are at least two reasons for this outcome. First, since a large change in the series is a rare event, there is not very much evidence in the subseries of 1000 observations about the nonlinearity. Here, the difference between RE of the estimated model and the corresponding measure for the known model was greater than in the previous case, and RE of the latter model remained below unity for all forecast horizons. Second, as argued in Clements and Hendry (1999), differencing helps construct models that adapt more quickly to large shifts in the series than models built on undifferenced data. This adaptability is demonstrated in the experiment of Lundbergh and Teräsvirta (2002). A very basic example emphasizing the same thing can be found in Hendry and Clements (2003).

These results also show that a model builder who begins his task by testing the unit root hypothesis may often end up with a model that is quite different from the one obtained by someone beginning by first testing linearity. In the present case, the latter course is perfectly defendable, because the data-generating process is stationary. The prevailing paradigm, testing the unit root hypothesis first, may thus not always be appropriate when the possibility of a nonlinear data-generating process cannot be excluded. For a discussion of the relationship between unit roots and nonlinearity; see Elliott (in press).

7 Empirical forecast comparisons

7.1 Relevant issues

The purpose of many empirical economic forecast comparisons involving nonlinear models is to find out whether, for a given time series or a set of series, nonlinear models yield more accurate forecasts than linear models. In many cases, the answer appears to be negative, even when the nonlinear model in question fits the data better than the corresponding linear model. Reasons for this outcome have been discussed in the literature. One argument put forward is that nonlinear models may sometimes explain features in the data that do not occur very frequently. If these features are not present in the series during the period to be forecast, then there is no gain from using nonlinear models for generating the forecasts. This may be the case at least when the number of out-of-sample forecasts is relatively small; see for example Teräsvirta and Anderson (1992) for discussion.

Essentially the same argument is that the nonlinear model can only be expected to forecast better than a linear one in particular regimes. For example, a nonlinear model may be useful in forecasting the volume of industrial production in recessions but not expansions. Montgomery et al. (1998) forecast the quarterly US unemployment rate using a two-regime threshold autoregressive model (7) and a two-regime Markov switching autoregressive model (8) . Both models, the SETAR model in particular, yield more accurate forecasts than the linear model when the forecasting origin lies in the recession. If it lies in the expansion, both models, now the MS-model in particular, perform clearly less well than the linear AR model. Considering Wolf's sunspot numbers, another nonlinear series, Tong and Moeanaddin (1988) showed that the values at the troughs of the sunspot cycle were forecast more accurately from a SETAR than from a linear model, whereas the reverse was true for the values around the peaks. An explanation to this finding may be that there is more variation over time in the height of the peaks than in the bottom value of the troughs.

Another potential reason for inferior performance of nonlinear models compared to linear ones is overfitting. A small example highlighting this possibility can be found in Granger and Teräsvirta (1991). The authors generated data from an STR model and fitted both a projection pursuit regression model (see Friedman and Stuetzle, 1981) and a linear model to the simulated series. When nonlinearity was strong (the error variance small), the projection pursuit approach led to more accurate forecasts than the linear model. When the evidence of nonlinearity was weak (the error variance large), the projection pursuit model overfitted, and the forecasts of the linear

model were more accurate than the ones produced by the projection pursuit model. Careful modelling, including testing linearity before fitting a nonlinear model as discussed in Section 3, reduces the likelihood of overfitting.

From the discussion in Section 6 it is also clear that in some cases, when the time series are short, having to estimate the parameters as opposed to knowing them will erase the edge that a correctly specified nonlinear model has compared to a linear approximation. Another possibility is that even if linearity is rejected when tested, the nonlinear model fitted to the time series is misspecified to the extent that its forecasting performance does not match the performance of a linear model containing the same variables. This situation is even more likely to occur if a nonlinear model nesting a linear one is fitted to the data without first testing linearity.

Finally, Dacco and Satchell (1999) showed that in regime-switching models, the possibility of misclassifying an observation when forecasting may lead to the forecasts on the average being inferior to the one from a linear model, although a regime-switching model known to the forecaster generates the data. The criterion for forecast accuracy is the mean squared forecast error. The authors give analytic conditions for this to be the case and do it using simple Markov-switching and SETAR models as examples.

7.2 Comparing linear and nonlinear models

Comparisons of the forecasting performance of linear and nonlinear models have often included only a limited number of models and time series. To take an example, Montgomery et al. (1998) considered forecasts of the quarterly US civilian employment series from a univariate Markov-switching model of type (8) and a SETAR model. They separated expansions and contractions from each other and concluded that SETAR and Markov-switching models are useful in forecasting recessions, whereas they do not perform better than linear models during expansions. Clements and Krolzig (1998) study the forecasts from the Markov-switching autoregressive model of type (10) and a threshold autoregressive model when the series to be forecast is the quarterly US gross national product. The main conclusion of their study was that nonlinear models do not forecast better than linear ones when the criterion is the RMSFE. Similar conclusions were reached by Siliverstovs and van Dijk (2003), Boero and Marrocu (2002) and Sarantis (1999) for a variety of nonlinear models and economic time series. Bradley and Jansen (2004) obtained this outcome for a US excess stock return series, whereas there was evidence that nonlinear models, including a STAR model, yield more accurate forecasts for industrial production than the linear autoregressive model. Kilian and Taylor (2003) concluded that in forecasting nominal exchange

rates, ESTAR models are superior to the random walk model, but only at long horizons, 2-3 years.

The RMSFE is a rather "academic" criterion for comparing forecasts. Granger and Pesaran (2000) emphasize the use of economic criteria that are based on the loss function of the forecaster. The loss function, in turn, is related to the decision problem at hand; for more discussion, see Granger and Machina (2006) . In such comparisons, forecasts from nonlinear models may fare better than in RMSFE comparisons. Satchell and Timmermann (1995) focussed on two loss functions: the MSFE and a payoff criterion based on the economic value of the forecast (forecasting the direction of change). When the MSFE increases, the probability of correctly forecasting the direction decreases if the forecast and the forecast error are independent. The authors showed that this need not be true when the forecast and the error are dependent of each other. They argued that this may often be the case for forecasts from nonlinear models.

Most forecast comparisons concern univariate or single-equation models. A recent exception is De Gooijer and Vidiella-i-Anguera (2004). The authors compared the forecasting performance of two bivariate threshold autoregressive models with cointegration with that of a linear bivariate vector error-correction model using two pairs of US macroeconomic series. For forecast comparisons, the RMSFE has to be generalized to the multivariate situation; see De Gooijer and Vidiella-i-Anguera (2004). The results indicated that the nonlinear models perform better than the linear one in an out-of-sample forecast exercise.

Some authors, including De Gooijer and Vidiella-i-Anguera (2004), have considered interval and density forecasts as well. The quality of such forecasts has typically been evaluated internally. For example, the assumed coverage probability of an interval forecast is compared to the observed coverage probability. This is a less than satisfactory approach when one wants to compare interval or density forecasts from different models. Corradi and Swanson (2006) survey tests developed for finding out which one of a set of misspecified models provides the most accurate interval or density forecasts. Since this is a very recent area of interest, there are hardly any applications yet of these tests to nonlinear models.

7.3 Large forecast comparisons

7.3.1 Forecasting with a separate model for each forecast horizon

As discussed in Section 4, there are two ways of constructing multiperiod forecasts. One may use a single model for each forecast horizon or construct

a separate model for each forecast horizon. In the former alternative, generating the forecasts may be computationally demanding if the number of variables to be forecast and the number of forecast horizons is large. In the latter, specifying and estimating the models may require a large amount of work, whereas forecasting is simple. In this section the focus is on a number of large studies that involve nonlinear models and several forecast horizons and in which separate models are constructed for each forecast horizon. Perhaps the most extensive such study is the one by Stock and Watson (1999). Other examples include Marcellino (2002) and Marcellino (2004). Stock and Watson (1999) forecast 215 monthly US macroeconomic variables, whereas Marcellino (2002) and Marcellino (2004) considered macroeconomic variables of the countries of the European Union.

The study of Stock and Watson (1999) involved two types of nonlinear models: a "tightly parameterized" model which was the LSTAR model of Section 2.3 and a "loosely parameterized" one, which was the autoregressive neural network model. The authors experimented with two families of AR-NN models: one with a single hidden layer, see (11) , and a more general family with two hidden layers. Various linear autoregressive models were included as well as models of exponential smoothing. Several methods of combining forecasts were included in comparisons. All told, the number of models or methods to forecast each series was 63.

The models were either completely specified in advance or the number of lags was specified using AIC or BIC. Two types of models were considered. Either the variables were in levels:

$$y_{t+h} = f_L(y_t, y_{t-1}, ..., y_{t-p+1}) + \varepsilon_t^L$$

where $h = 1, 6$ or 12, or they were in differences:

$$y_{t+h} - y_t = f_D(\Delta y_t, \Delta y_{t-1}, ..., \Delta y_{t-p+1}) + \varepsilon_t^D.$$

The experiment incuded several values of p. The series were forecast every month starting after a startup period of 120 observations. The last observation in all series was 1996(12), and for most series the first observation was 1959(1). The models were re-estimated and, in the case of combined forecasts, the weights of the individual models recalculated every month. The insanity filter that the authors called trimming of forecasts was applied. The purpose of the filter was to make the process better mimic the behaviour of a true forecaster.

The 215 time series covered most types of macroeconomic series from production, consumption, money and credit series to stock returns. The series that originally contained seasonality were seasonally adjusted.

The forecasting methods were ranked according to several criteria. A general conclusion was that the nonlinear models did not perform better than the linear ones. In one comparison, the 63 different models and methods were ranked on forecast performance using three different loss functions, the absolute forecast errors raised to the power one, two, or three, and the three forecast horizons. The best ANN forecast had rank around 10, whereas the best STAR model typically had rank around 20. The combined forecasts topped all rankings, and, interestingly, combined forecasts of nonlinear models only were always ranked one or two. The best linear models were better than the STAR models and, at longer horizons than one month, better than the ANN models. The no-change model was ranked among the bottom two in all rankings showing that all models had at least some relevance as forecasting tools.

A remarkable result, already evident from the previous comments, was that combining the forecasts from all nonlinear models generated forecasts that were among the most accurate in rankings. They were among the top five in 53% (models in levels) and 51% (models in differences) of all cases when forecasting one month ahead. This was by far the highest fraction of all methods compared. In forecasting six and twelve months ahead, these percentages were lower but still between 30% and 34%. At these horizons, the combinations involving all linear models had a comparable performance. All single models were left far behind. Thus a general conclusion from the study of Stock and Watson is that there is some exploitable nonlinearity in the series under consideration, but that it is too diffuse to be captured by a single nonlinear model.

Marcellino (2002) reported results on forecasting 480 variables representing the economies of the twelve countries of the European Monetary Union. The monthly time series were shorter than the series in Stock and Watson (1999), which was compensated for by a greater number of series. There were 58 models but, unlike Stock and Watson, Marcellino did not consider combining forecasts from them. In addition to linear models, neural network models and logistic STAR models were included in the study. A novelty, compared to Stock and Watson (1999), was that a set of time-varying autoregressive models of type (15) was included in the comparisons.

The results were based on rankings of models' performance measured using loss functions based on absolute forecast errors now raised to five powers from one to three in steps of 0.5. Neither neural network nor LSTAR models appeared in the overall top-10. But then, both the fraction of neural network models and LSTAR models that appeared in top-10 rankings for individual series was greater than the same fraction for linear methods or time-varying AR models. This, together with other results in the paper, suggests that

nonlinear models in many cases work very well, but they can also relatively often perform rather poorly.

Marcellino (2002) also singled out three 'key economic variables': the growth rate of industrial production, the unemployment rate and the inflation measured by the consumer price index. Ranking models within these three categories showed that industrial production was best forecast by linear models. But then, in forecasting the unemployment rate, both the LSTAR and neural network models, as well as the time-varying AR model, had top rankings. For example, for the three-month horizon, two LSTAR models occupied the one-two ranks for all five loss functions (other ranks were not reported). This may not be completely surprising since many European unemployment rate series are distinctly asymmetric; see for example Skalin and Teräsvirta (2002) for discussion based on quarterly series. As to the inflation rate, the results were a mixture of the ones for the other two key variables.

These studies suggest some answers to the question of whether nonlinear models perform better than linear ones in forecasting macroeconomic series. The results in Stock and Watson (1999) indicate that using a large number of nonlinear models and combining forecasts from them is much better than using single nonlinear models. It also seems that this way of exploiting nonlinearity may lead to better forecasting performance than what is achieved by linear models. Marcellino (2002) did not consider this possibility. His results, based on individual models, suggest that nonlinear models are uneven performers but that they can do well in some types of macroeconomic series such as unemployment rates.

7.3.2 Forecasting with the same model for each forecast horizon

As discussed in Section 4, it is possible to obtain forecasts for several periods ahead recursively from a single model. This is the approach adopted in Teräsvirta et al. (2005). The main question posed in that paper was whether careful modelling improves forecast accuracy compared to models with a fixed specification that remains unchanged over time. In the case of nonlinear models this implied testing linearity first and choosing a nonlinear model only if linearity is rejected. The lag structure of the nonlinear model was also determined from the data. The authors considered seven monthly macroeconomic variables of the G7 countries. They were industrial production, unemployment, volume of exports, volume of imports, inflation, narrow money, and short-term interest rate. Most series started in January 1960 and were available up to December 2000. The series were seasonally adjusted with the exception of the CPI inflation and the short-term interest rate. As in Stock and Watson (1999), the series were forecast every month.

In order to keep the human effort and computational burdens at manageable levels, the models were only respecified every 12 months.

The models considered were the linear autoregressive model, the LSTAR model and the single hidden-layer feedforward neural network model. The results showed that there were series for which linearity was never rejected. Rejections, using LM-type tests, were somewhat more frequent against LSTAR than against the neural network model. The interest rate series, the inflation rate and the unemployment rate were most systematically nonlinear when linearity was tested against STAR. In order to find out whether modelling was a useful idea, the investigation also included a set of models with a predetermined form and lag structure.

Results were reported for four forecast horizons: 1, 3, 6 and 12 months. They indicated that careful modelling does improve the accuracy of forecasts compared to selecting fixed nonlinear models. The loss function was the root mean square error. The LSTAR model turned out to be the best model overall, better than the linear or neural network model, which was not the case in Stock and Watson (1999) or Marcellino (2002). The LSTAR model did not, however, dominate the others. There were series/country pairs for which other models performed clearly better than the STAR model. Nevertheless, as in Marcellino (2002), the LSTAR model did well in forecasting the unemployment rate.

The results on neural network models suggested the need for model evaluation: a closer scrutiny found some of the estimated models to be explosive, which led to inferior multi-step forecasts. This fact emphasizes the need for model evaluation before forecasting. For practical reasons, this phase of model building has been neglected in large studies such as the ones discussed in this section.

The results in Teräsvirta et al. (2005) are not directly comparable to the ones in Stock and Watson (1999) or Marcellino (2002) because the forecasts in the former paper have been generated recursively from a single model for all forecast horizons. The time series used in these three papers have not been the same either. Nevertheless, put together the results strengthen the view that nonlinear models are a useful tool in macroeconomic forecasting.

8 Final remarks

This chapter contains a presentation of a number of frequently applied nonlinear models and shows how forecasts can be generated from them. Since such forecasts are typically obtained numerically when the same model is used for forecasting several periods ahead, forecast generation automatically

yields not only point but interval and density forecasts as well. The latter are important because they contain more information than the pure point forecasts which, unfortunately, often are the only ones reported in publications. It is also sometimes argued that the strength of the nonlinear forecasting lies in density forecasts, whereas comparisons of point forecasts often show no substantial difference in performance between individual linear and nonlinear models. Results from large studies reported in Section 7.3 indicate that forecasts from linear models may be more robust than the ones from nonlinear models. In some cases the nonlinear models clearly outperform the linear ones, but in other occasions they may be strongly inferior to the latter.

It appears that nonlinear models may have a fair chance of generating accurate forecasts if the number of observations for specifying the model and estimating its parameters is large. This is due to the fact, discussed in Lundbergh and Teräsvirta (2002), that potential gains from forecasting with nonlinear models can be strongly reduced because of parameter estimation. A recent simulation-based paper by Psaradakis and Spagnolo (2005), where the observations are generated by a bivariate nonlinear system, either a threshold model or a Markov-switching one, with linear cointegration, strengthens this impression. In some cases, even when the data-generating process is nonlinear and the model is correctly specified, the linear model yields more accurate forecasts than the correct nonlinear one with estimated parameters. Short time series are thus a disadvantage, but the results also suggest that sufficient attention should be paid to estimation techniques. This is certainly true for neural network models that contain a large number of parameters. Recent developments in this area include White (2006).

In the nonlinear framework, the question of iterative vs. direct forecasts requires more research. Simulations reported in Lin and Granger (1994) suggest that the direct method is not a useful alternative when the data-generating process is a nonlinear model such as the STAR model, and a direct STAR model is fitted to the data for forecasting more than one period ahead. The direct method works better when the model used to produce the forecasts is a neural network model. This may not be surprising because the neural network model is a flexible functional form. Whether direct nonlinear models generate more accurate forecasts than direct linear ones when the data-generating process is nonlinear, is a topic for further research.

An encouraging feature is, however, that there is evidence of combination of a large number of nonlinear models leading to point forecasts that are superior to forecasts from linear models. Thus it may be concluded that while the form of nonlinearity in macroeconomic time series may be difficult to usefully capture with single models, there is hope for improving forecasting accuracy by combining information from several nonlinear models. This sug-

gests that parametric nonlinear models will remain important in forecasting economic variables.

References

Aiolfi, M. and Timmermann, A.: in press, Persistence in forecasting performance and conditional combination strategies, *Journal of Econometrics* .

Andersen, T., Bollerslev, T. and Christoffersen, P.: 2006, Volatility forecasting, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Andrews, D. W. K. and Ploberger, W.: 1994, Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica* **62**, 1383–1414.

Bacon, D. W. and Watts, D. G.: 1971, Estimating the transition between two intersecting straight lines, *Biometrika* **58**, 525–534.

Bai, J. and Perron, P.: 1998, Estimating and testing linear models with multiple structural changes, *Econometrica* **66**, 47–78.

Bai, J. and Perron, P.: 2003, Computation and analysis of multiple structural change models, *Journal of Applied Econometrics* **18**, 1–22.

Banerjee, A. and Urga, G.: 2005, Modelling structural breaks, long memory and stock market volatility: An overview, *Journal of Econometrics* **129**, 1–34.

Bhansali, R. J.: 2002, Multi-step forecasting, *in* M. P. Clements and D. F. Hendry (eds), *A Companion to Economic Forecasting*, Blackwell, Oxford, pp. 206–221.

Bierens, H. J.: 1990, A consistent conditional moment test of functional form, *Econometrica* **58**, 1443–1458.

Boero, G. and Marrocu, E.: 2002, The performance of non-linear exchange rate models: A forecasting comparison, *Journal of Forecasting* **21**, 513–542.

Box, G. E. P. and Jenkins, G. M.: 1970, *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco.

Bradley, M. D. and Jansen, D. W.: 2004, Forecasting with a nonlinear dynamic model of stock returns and industrial production, *International Journal of Forecasting* **20**, 321–342.

Brännäs, K. and De Gooijer, J. G.: 1994, Autoregressive - asymmetric moving average model for business cycle data, *Journal of Forecasting* **13**, 529–544.

Breunig, R., Najarian, S. and Pagan, A.: 2003, Specification testing of Markov switching models, *Oxford Bulletin of Economics and Statistics* **65**, 703–725.

Brown, B. W. and Mariano, R. S.: 1984, Residual-based procedures for prediction and estimation in a nonlinear simultaneous system, *Econometrica* **52**, 321–343.

Chan, K. S.: 1993, Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model, *Annals of Statistics* **21**, 520–533.

Chan, K. S. and Tong, H.: 1986, On estimating thresholds in autoregressive models, *Journal of Time Series Analysis* **7**, 178–190.

Clements, M. P., Franses, P. H. and Swanson, N. R.: 2004, Forecasting economic and financial time-series with non-linear models, *International Journal of Forecasting* **20**, 169–183.

Clements, M. P. and Hendry, D. F.: 1999, *Forecasting Non-stationary Economic Time Series*, MIT Press, Cambridge, MA.

Clements, M. P. and Krolzig, H.-M.: 1998, A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP, *Econometrics Journal* **1**, C47–C75.

Corradi, V. and Swanson, N. R.: 2002, A consistent test for non-linear out of sample predictive accuracy, *Journal of Econometrics* **110**, 353–381.

Corradi, V. and Swanson, N. R.: 2004, Some recent developments in predictive accuracy testing with nested models and (generic) nonlinear alternatives, *International Journal of Forecasting* **20**, 185–199.

Corradi, V. and Swanson, N. R.: 2006, Predictive density evaluation, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Cybenko, G.: 1989, Approximation by superposition of sigmoidal functions, *Mathematics of Control, Signals, and Systems* **2**, 303–314.

Dacco, R. and Satchell, S.: 1999, Why do regime-switching models forecast so badly?, *Journal of Forecasting* **18**, 1–16.

Davies, R. B.: 1977, Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* **64**, 247–254.

De Gooijer, J. G. and De Bruin, P. T.: 1998, On forecasting SETAR processes, *Statistics and Probability Letters* **37**, 7–14.

De Gooijer, J. G. and Vidiella-i-Anguera, A.: 2004, Forecasting threshold cointegrated systems, *International Journal of Forecasting* **20**, 237–253.

Deutsch, M., Granger, C. W. J. and Teräsvirta, T.: 1994, The combination of forecasts using changing weights, *International Journal of Forecasting* **10**, 47–57.

Diebold, F. X. and Mariano, R. S.: 1995, Comparing predictive accuracy, *Journal of Business and Economic Statistics* **13**, 253–263.

Eitrheim, Ø. and Teräsvirta, T.: 1996, Testing the adequacy of smooth transition autoregressive models, *Journal of Econometrics* **74**, 59–75.

Elliott, G.: in press, Forecasting with trending data, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Enders, W. and Granger, C. W. J.: 1998, Unit-root tests and asymmetric adjustment with an example using the term structure of interest rates, *Journal of Business and Economic Statistics* **16**, 304–311.

Fan, J. and Yao, Q.: 2003, *Nonlinear Time Series. Nonparametric and Parametric Methods*, Springer, New York.

Fine, T. L.: 1999, *Feedforward Neural Network Methodology*, Springer-Verlag, Berlin.

Franses, P. H. and van Dijk, D.: 2000, *Non-Linear Time Series Models in Empirical Finance*, Cambridge University Press, Cambridge.

Friedman, J. H. and Stuetzle, W.: 1981, Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817–823.

Funahashi, K.: 1989, On the approximate realization of continuous mappings by neural networks, *Neural Networks* **2**, 183–192.

Garcia, R.: 1998, Asymptotic null distribution of the likelihood ratio test in Markov switching models, *International Economic Review* **39**, 763–788.

Giacomini, R. and White, H.: 2003, Tests of conditional predictive ability, *Working paper 2003-09*, Department of Economics, University of California, San Diego.

Goffe, W. L., Ferrier, G. D. and Rogers, J.: 1994, Global optimization of statistical functions with simulated annealing, *Journal of Econometrics* **60**, 65–99.

Gonzalo, J. and Pitarakis, J.-Y.: 2002, Estimation and model selection based inference in single and multiple threshold models, *Journal of Econometrics* **110**, 319–352.

Granger, C. W. J. and Bates, J.: 1969, The combination of forecasts, *Operations Research Quarterly* **20**, 451–468.

Granger, C. W. J. and Jeon, Y.: 2004, Thick modeling, *Economic Modelling* **21**, 323–343.

Granger, C. W. J. and Machina, M. J.: 2006, Forecasting and decision theory, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Granger, C. W. J. and Pesaran, M. H.: 2000, Economic and statistical measures of forecast accuracy, *Journal of Forecasting* **19**, 537–560.

Granger, C. W. J. and Teräsvirta, T.: 1991, Experiments in modeling nonlinear relationships between time series, *in* M. Casdagli and S. Eubank (eds), *Nonlinear Modeling and Forecasting*, Addison-Wesley, Redwood City, pp. 189–197.

Granger, C. W. J. and Teräsvirta, T.: 1993, *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

Haggan, V. and Ozaki, T.: 1981, Modelling non-linear random vibrations using an amplitude-dependent autoregressive time series model, *Biometrika* **68**, 189–196.

Hamilton, J. D.: 1989, A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* **57**, 357–384.

Hamilton, J. D.: 1993, Estimation, inference and forecasting of time series subject to changes in regime, *in* G. S. Maddala, C. R. Rao and H. R. Vinod (eds), *Handbook of Statistics*, Vol. 11, Elsevier, Amsterdam, pp. 231–260.

Hamilton, J. D.: 1994, *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Hamilton, J. D.: 1996, Specification testing in Markov-switching time-series models, *Journal of Econometrics* **70**, 127–157.

Hansen, B. E.: 1996, Inference when a nuisance parameter is not identified under the null hypothesis, *Econometrica* **64**, 413–430.

Hansen, B. E.: 1999, Testing for linearity, *Journal of Economic Surveys* **13**, 551–576.

Harvey, A. C.: 2006, Forecasting with unobserved components time series models, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Harvey, D., Leybourne, S. and Newbold, P.: 1997, Testing the equality of prediction mean squared errors, *International Journal of Forecasting* **13**, 281–291.

Haykin, S.: 1999, *Neural Networks. A Comprehensive Foundation*, second edn, Prentice Hall, Upper Saddle River, NJ.

Hendry, D. F. and Clements, M. P.: 2003, Economic forecasting: Some lessons from recent research, *Economic Modelling* **20**, 301–329.

Henry, O. T., Olekalns, N. and Summers, P. M.: 2001, Exchange rate instability: A threshold autoregressive approach, *Economic Record* **77**, 160–166.

Hornik, K., Stinchombe, M. and White, H.: 1989, Multi-layer Feedforward networks are universal approximators, *Neural Networks* **2**, 359–366.

Hwang, J. T. G. and Ding, A. A.: 1997, Prediction intervals for artificial neural networks, *Journal of the American Statistical Association* **92**, 109–125.

Hyndman, R. J.: 1996, Computing and graphing highest density regions, *The American Statistician* **50**, 120–126.

Inoue, A. and Kilian, L.: 2004, In-sample or out-of-sample tests of predictability: Which one should we use?, *Econometric Reviews* **23**, 371–402.

Kilian, L. and Taylor, M. P.: 2003, Why is it so difficult to beat the random walk forecast of exchange rates?, *Journal of International Economics* **60**, 85–107.

Lee, T.-H., White, H. and Granger, C. W. J.: 1993, Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests, *Journal of Econometrics* **56**, 269–290.

Li, H. and Xu, Y.: 2002, Short rate dynamics and regime shifts, *Working paper*, Johnson Graduate School of Management, Cornell University.

Lin, C.-F. and Teräsvirta, T.: 1999, Testing parameter constancy in linear models against stochastic stationary parameters, *Journal of Econometrics* **90**, 193–213.

Lin, J.-L. and Granger, C. W. J.: 1994, Forecasting from non-linear models in practice, *Journal of Forecasting* **13**, 1–9.

Lindgren, G.: 1978, Markov regime models for mixed distributions and switching regressions, *Scandinavian Journal of Statistics* **5**, 81–91.

Lundbergh, S. and Teräsvirta, T.: 2002, Forecasting with smooth transition autoregressive models, *in* M. P. Clements and D. F. Hendry (eds), *A Companion to Economic Forecasting*, Blackwell, Oxford, pp. 485–509.

Luukkonen, R., Saikkonen, P. and Teräsvirta, T.: 1988, Testing linearity against smooth transition autoregressive models, *Biometrika* **75**, 491–499.

Maddala, D. S.: 1977, *Econometrics*, McGraw-Hill, New York.

Marcellino, M.: 2002, Instability and non-linearity in the EMU, *Discussion Paper No. 3312*, Centre for Economic Policy Research.

Marcellino, M.: 2004, Forecasting EMU macroeconomic variables, *International Journal of Forecasting* **20**, 359–372.

Marcellino, M., Stock, J. H. and Watson, M. W.: 2004, A comparison of direct and iterated multistep AR methods for forecasting economic time series, *Working paper*.

Medeiros, M. C., Teräsvirta, T. and Rech, G.: in press, Building neural network models for time series: A statistical approach, *Journal of Forecasting* .

Mincer, J. and Zarnowitz, V.: 1969, The evaluation of economic forecasts, *in* J. Mincer (ed.), *Economic Forecasts and Expectations*, National Bureau of Economic Research, New York.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S. and Tiao, G. C.: 1998, Forecasting the U.S. unemployment rate, *Journal of the American Statistical Association* **93**, 478–493.

Nyblom, J.: 1989, Testing for the constancy of parameters over time, *Journal of the American Statistical Association* **84**, 223–230.

Pesaran, M. H. and Timmermann, A.: 2002, Model instability and choice of observation window, *Working paper*.

Pfann, G. A., Schotman, P. C. and Tschernig, R.: 1996, Nonlinear interest rate dynamics and implications for term structure, *Journal of Econometrics* **74**, 149–176.

Poon, S. H. and Granger, C. W. J.: 2003, Forecasting volatility in financial markets, *Journal of Economic Literature* **41**, 478–539.

Proietti, T.: 2003, Forecasting the US unemployment rate, *Computational Statistics and Data Analysis* **42**, 451–476.

Psaradakis, Z. and Spagnolo, F.: 2005, Forecast performance of nonlinear error-correction models with multiple regimes, *Journal of Forecasting* **24**, 119–138.

Ramsey, J. B.: 1996, If nonlinear models cannot forecast, what use are they?, *Studies in Nonlinear Dynamics and Forecasting* **1**, 65–86.

Sarantis, N.: 1999, Modelling non-linearities in real effective exchange rates, *Journal of International Money and Finance* **18**, 27–45.

Satchell, S. and Timmermann, A.: 1995, An assessment of the economic value of non-linear foreign exchange rate forecasts, *Journal of Forecasting* **14**, 477–497.

Siliverstovs, B. and van Dijk, D.: 2003, Forecasting industrial production with linear, nonlinear, and structural change models, *Econometric Institute Report EI 2003-16*, Erasmus University Rotterdam.

Skalin, J. and Teräsvirta, T.: 2002, Modeling asymmetries and moving equilibria in unemployment rates, *Macroeconomic Dynamics* **6**, 202–241.

Stock, J. H. and Watson, M. W.: 1999, A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, *in* R. F. Engle and H. White (eds), *Cointegration, Causality and Forecasting. A Festschrift in Honour of Clive W.J. Granger*, Oxford University Press, Oxford, pp. 1–44.

Strikholm, B. and Teräsvirta, T.: 2005, Determining the number of regimes in a threshold autoregressive model using smooth transition autoregressions, *Working Paper 578*, Stockholm School of Economics.

Swanson, N. R. and White, H.: 1995, A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks, *Journal of Business and Economic Statistics* **13**, 265–275.

Swanson, N. R. and White, H.: 1997a, Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models, *International Journal of Forecasting* **13**, 439–461.

Swanson, N. R. and White, H.: 1997b, A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks, *Review of Economic and Statistics* **79**, 540–550.

Tay, A. S. and Wallis, K. F.: 2002, Density forecasting: A survey, *in* M. P. Clements and D. F. Hendry (eds), *A Companion to Economic Forecasting*, Blackwell, Oxford, pp. 45–68.

Taylor, M. P. and Sarno, L.: 2002, Purchasing power parity and the real exchange rate, *International Monetary Fund Staff Papers* **49**, 65–105.

Teräsvirta, T.: 1994, Specification, estimation, and evaluation of smooth transition autoregressive models, *Journal of the American Statistical Association* **89**, 208–218.

Teräsvirta, T.: 1998, Modeling economic relationships with smooth transition regressions, *in* A. Ullah and D. E. Giles (eds), *Handbook of Applied Economic Statistics*, Dekker, New York, pp. 507–552.

Teräsvirta, T.: 2004, Nonlinear smooth transition modeling, *in* H. Lütkepohl and M. Krätzig (eds), *Applied Time Series Econometrics*, Cambridge University Press, Cambridge, pp. 222–242.

Teräsvirta, T. and Anderson, H. M.: 1992, Characterizing nonlinearities in business cycles using smooth transition autoregressive models, *Journal of Applied Econometrics* **7**, S119–S136.

Teräsvirta, T. and Eliasson, A.-C.: 2001, Non-linear error correction and the UK demand for broad money, 1878-1993, *Journal of Applied Econometrics* **16**, 277–288.

Teräsvirta, T., Lin, C.-F. and Granger, C. W. J.: 1993, Power of the neural network linearity test, *Journal of Time Series Analysis* **14**, 309–323.

Teräsvirta, T., van Dijk, D. and Medeiros, M. C.: 2005, Smooth transition autoregressions, neural networks, and linear models in forecasting macroeconomic time series: A re-examination, *International Journal of Forecasting* **21**, 755–774.

Timmermann, A.: 2006, Forecast combinations, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Tong, H.: 1990, *Non-Linear Time Series. A Dynamical System Approach*, Oxford University Press, Oxford.

Tong, H. and Moeanaddin, R.: 1988, On multi-step nonlinear least squares prediction, *The Statistician* **37**, 101–110.

Tsay, R. S.: 2002, Nonlinear models and forecasting, *in* M. P. Clements and D. F. Hendry (eds), *A Companion to Economic Forecasting*, Blackwell, Oxford, pp. 453–484.

Tyssedal, J. S. and Tjøstheim, D.: 1988, An autoregressive model with suddenly changing parameters, *Applied Statistics* **37**, 353–369.

van Dijk, D., Teräsvirta, T. and Franses, P. H.: 2002, Smooth transition autoregressive models - a survey of recent developments, *Econometric Reviews* **21**, 1–47.

Venetis, I. A., Paya, I. and Peel, D. A.: 2003, Re-examination of the predictability of economic activity using the yield spread: A nonlinear approach, *International Review of Economics and Finance* **12**, 187–206.

Wallis, K. F.: 1999, Asymmetric density forecasts of inflation and the Bank of England's fan chart, *National Institute Economic Review* **167**, 106–112.

Watson, M. W. and Engle, R. F.: 1985, Testing for regression coefficient stability with a stationary AR(1) alternative, *Review of Economics and Statistics* **67**, 341–346.

Wecker, W. E.: 1981, Asymmetric time series, *Journal of the American Statistical Association* **76**, 16–21.

West, K. D.: 2006, Forecast evaluation, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

White, H.: 1990, Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings, *Neural Networks* **3**, 535–550.

White, H.: 2006, Approximate nonlinear forecasting methods, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Zhang, G., Patuwo, B. E. and Hu, M. Y.: 1998, Forecasting with artificial neural networks: The state of the art, *International Journal of Forecasting* **14**, 35–62.